Frederico Miguel da Conceição Lourenço

Graduated in Biochemistry

Assignment of new roles for malectin-like domains to understand their divergent evolution

Dissertation to obtain Master degree in Biochemistry

Supervisor: Benedita Andrade Pinheiro, UCIBIO, FCT/UNL

Jury:

Examiner: Doutor Jorge da Silva Dias
Vowel: Doutora Benedita Andrade Pinheiro

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

October 2017

Frederico Miguel da Conceição Lourenço
Graduated in Biochemistry

Assignment of new roles for malectin-like domains to understand their divergent evolution

Dissertation to obtain Master degree in
Biochemistry

Supervisor: Benedita Andrade Pinheiro, UCIBIO, FCT/UNL

Jury:

Examiner: Doutor Jorge da Silva Dias
Vowel: Doutora Benedita Andrade Pinheiro

October 2017

“Assignment of new roles for malectin-like domains to understand their divergent evolution”

Acknowledgments:

À minha orientadora, **Doutora Benedita Pinheiro**, por acreditar em mim e no meu trabalho. Tive muita sorte em ter como orientadora, pelo apoio e pela aprendizagem.

À **Professora Doutora Maria João Romão**, não apenas como líder do grupo de investigação XTAL da FCT-UNL, mas onde me recebeu e pela oportunidade de realizar o meu trabalho no Laboratório de Cristalografia de proteínas.

À **Doutora Angelina Palma** pelo ensinamento e orientação na técnica de microarrays de glicanos.

À **Doutora Ana Luísa** pelo acompanhamento da difração de cristais *in house*.

À **Viviana Correia** que me apoiou e ajudou em toda a parte laboratorial, bem como sempre disponível quando me surgia alguma dúvida.

À **Professora Ten Feizi**, ao **Doutor Wengand Chai** e à **Doutora Lisete Silva** do Glycosciences Laboratory, Imperial College London pela colaboração na construção do robótico microarray de glicanos.

Ao **Professor Manuel Coimbra**, **Doutora Cláudia Nunes** e **Carolina Pandeirada**, da Universidade de Aveiro, e à **Professora Smestad Paulsen Berit** da Universidade de Oslo pelo fornecimento de amostras de polissacáridos para a construção dos microarrays.

A todos os colegas do grupo **XTAL**, pelo fantástico ambiente de trabalho, pela simpatia e sempre disponíveis para ajudar em qualquer questão.

À **Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa**, uma faculdade que me proporcionou toda a minha formação académica, e também a todos os **docentes** de licenciatura e mestrado em Bioquímica por todos os ensinamentos adquiridos.

Ao **BAG Portugal** pela oportunidade de recolher os dados de difração de cristais no European Synchrotron Radiation Facility (ESRF), em Grenoble.

À **Fundação para a Ciência e Tecnologia (FCT-MCTES)** por ter financiado o trabalho através dos projectos PTDC/BBB-BEP/0869/2014, PTDC/QUI-QUI/112537/2009 e RECI/BBB-BEP/0124/2012. A unidade de investigação UCIBIO (Unidade de Ciências e Biomoleculares Aplicadas) é financiada por fundos nacionais da FCT-MCTES (UID/Multi/04378/2013) e co-financiada pelo FEDER no âmbito do PT2020 (POCI-01-0145-FEDER-007728)

Um grande obrigado aos meus **pais**, por nunca terem duvidado de mim. Sem vocês não estava onde cheguei.

Um obrigado aos meus **avós**, adoro-vos.

Abstract:

Malectin is a highly-conserved animal lectin from the endoplasmic reticulum (ER), with a quality control function in the *N*-Glycosylation process. It has a β-sandwich core with long loops connecting the β-sheets. Malectin binding-pocket is in the loops region. Several carbohydrate-binding modules (CBMs) discovered in other domains of life that shared sequence homology with the malectin, were classified and grouped as a novel CBM57 family by Carbohydrate-Active Enzymes (CAZy) database. The members of this family are expected to have a highly conserved β-sandwich core, but high variance in the binding-pocket residues.

To investigate if the specificity of these modules is the same as the malectin, a bioinformatic analysis was performed with 315 members of the CBM57 family found in CAZy database. Several programs were used to predict the protein architecture and to analyse the conservation of amino acids sequences, especially in the binding-pocket. Based on this analysis, we predict animal CBM57 modules to have the same specificity as malectin. However, bacterial CBM57 modules in bacteria domain are predicted, after highlighting the modules associated with glycoside hydrolases from family 2, to have various specificities, and thus different biological functions. For verifying these assumptions, a total of 7 CBMs (family 57 and homologous) associated with glycoside hydrolases from family 2 and belonging to the human gut microbiome – *Bacteroides ovatus* and *Bacteroides thetaiotaomicron*- were chosen for characterization studies.

A re-cloning was initially performed for the recombinant DNAs, changing the His-tag position. Afterwards, expression tests were realized, in which 2 CBMs of different bacteria were expressed in soluble form. The production of the proteins was then performed at a larger scale, followed by affinity chromatography purification. By the analysis of the gels, the eluted samples had high purity and were suitable for characterization studies.

Glycan microarrays were performed for determining the binding-specificities of the 2 CBM modules. The CBM module from *B.thetaiotaomicron* revealed high specificity for pectin polysaccharides, possible recognizing α 1-3 linked galacturonic acid and ramnose. For structural characterization by X-ray crystallography, several crystallization trials were performed. Crystals were obtained for the *B.thetaiotaomicron* CBM module, which diffracted to high resolution. The structure is, yet, to be solved.

Keywords: Malectin, CBM, bioinformatic analysis, glycan microarray, X-ray crystallography

Resumo:

A malectina é uma lectina do retículo endoplasmático (ER), muito conservada em animais e com função de controlo de qualidade dos processos de *N*-glicosilação. A malectina tem uma estrutura β-sandwich composta por folhas-β ligadas por longos *loops* que formam um *pocket* de ligação. Vários módulos de ligação a hidratos de carbono (CBMs) descobertos noutros Domínios da Vida partilham homologia de sequência com a da malectina, sendo classificados e agrupados na base de dados Carbohydrate-Active Enzymes (CAZy) como uma nova família CBM57. Nesta família é expectável os módulos terem a estrutura β-sandwich conservada, mas com diferenças ao nível dos resíduos de ligação a açúcares.

Para investigar possíveis diferenças na especificidade destes módulos, uma análise bioinformática foi realizada para os 315 módulos de CBM57 encontrados na base de dados CAZy, usando vários programas para a previsão da arquitectura das proteínas e analise da conservação da sequência de aminoácidos. Com base nesta análise, é expectável a conservação de especificidade nos dominio CBM57 de animais. No entanto é previsto que domínios CBM57 de bactérias, principalmente quando associados a hidrolases glicosidicas da família 2, terem especificidades diferentes entre eles. Para averiguar tal hipótese, um total de 7 CBMs (membros da família 57 e homólogos) associadas a hidrolases glicosidicas da familia 2, encontradas em duas espécies de bactérias pertencendo ao microbioma humano, foram escolhidas para a caracterização bioquimica.

Uma re-clonagem inicial foi feita dos DNA recombinantes para a troca da posição da cauda de histidinas. Posteriormente, testes de expressão foram realizados, tendo 2 CBMs de duas bactérias foram expressos na forma solúvel. Feitos os crescimentos em maior escala destes 2 CBMs, foi feita a purificação por cromatografia de afinidade. Através da análise de géis, as concluiu-se que as amostras eluidas apresentavam uma grande pureza e suficiente para estudos de caracterização.

Para a determinação da especificidade de ligação dos 2 módulos CBM foram realizados microarrays de glicanos. O módulo CBM do *B.thetaiotaomicron* revelou alta especificidade para pectinas, sendo os resíduos ácido galacturónico e ramnose ligados por uma ligação glicosíca α 1-3 o possivel epitopo reconhecido. Para a caracterização estrutural por cristalografia de raios-X, vários ensaios de cristalização foram realizados. Para o modulo CBM do *B. thetaiotaomicron*, foram obtidos cristais que difrataram a uma alta resolução. No entanto, a sua estrutura está por resolver.

Palavras-chaves: Malectina, CBM, análise bioinformática, microarray de glicanos, cristalografia de raios-.X

Contest:

List of figures:

List of tables:

Abbreviations and Symbols:

% (w/v)- weight/volume percentage

Å- Angstrom

B. ovatus- Bacteroides ovatus

B. thetaiotaomicron- Bacteroides *thetaiotaomicron*

CAZymes- Carbohydrate active enzymes

CBM- Carbohydrate-binding module

E. coli- Escherichia coli

ERSF- European Synchrotron Radiation Facility

GH2- Glycoside hydrolase from family 2

HEPES- 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid

IMAC- Immobilized metal affinity chromatography

IPTG- Isopropyl β-D-1-thiogalactopyranoside

KDa- Kilodalton

LB- Luria-Bertani

NMR- Nuclear Magnetic Resonance

MAD-Multiple Wavelength Anomalous Dispersion

MIR- Multiple Isomorphous Replacement

MR- Molecular Replacement

O.D.600nm- Optical density at 600 nm

PDB- Protein Data Bank

PKD- Polycystic Kidney disease

RCA I- *Ricinus Communis Agglutinin I*

PEG - Polyethylene glycol

PGA- Polygalacturonate acid

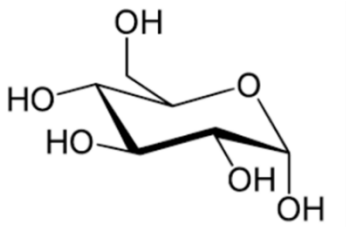
RG I- Rhamnogalacturonan I

RG II- Rhamnogalacturonan II

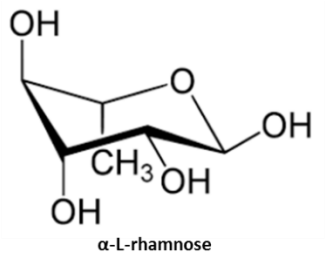
Rpm– Rotations per minute

SAD- Single Wavelength Anomalous Dispersion

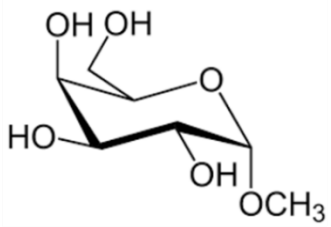
Chemical structures of monosaccharides present in glycan probes presented in this thesis



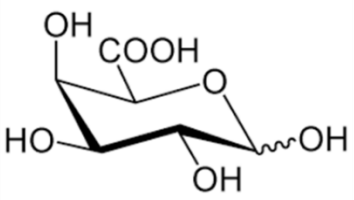
α-D-glucose



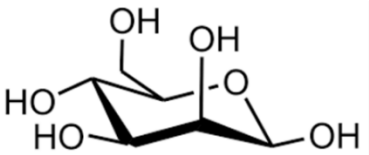
α-L-rhamnose



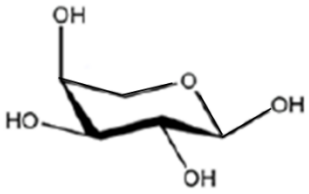
α-D-galactose



α/β-D-galacturonic acid



β-D-mannose



α-L-arabinopyranose

The symbol notation for monosaccharides is in accord with the proposed in the Glycosciences Laboratory, Imperial College (https://glycosciences.med.ic.ac.uk/docs/symbols.pdf).

Throughout this thesis, oligo- and polysaccharide structures were written in condensed or short forms according to the most recent version of carbohydrates nomenclature, established by International Union of Pure and Applied Chemistry (IUPAC).

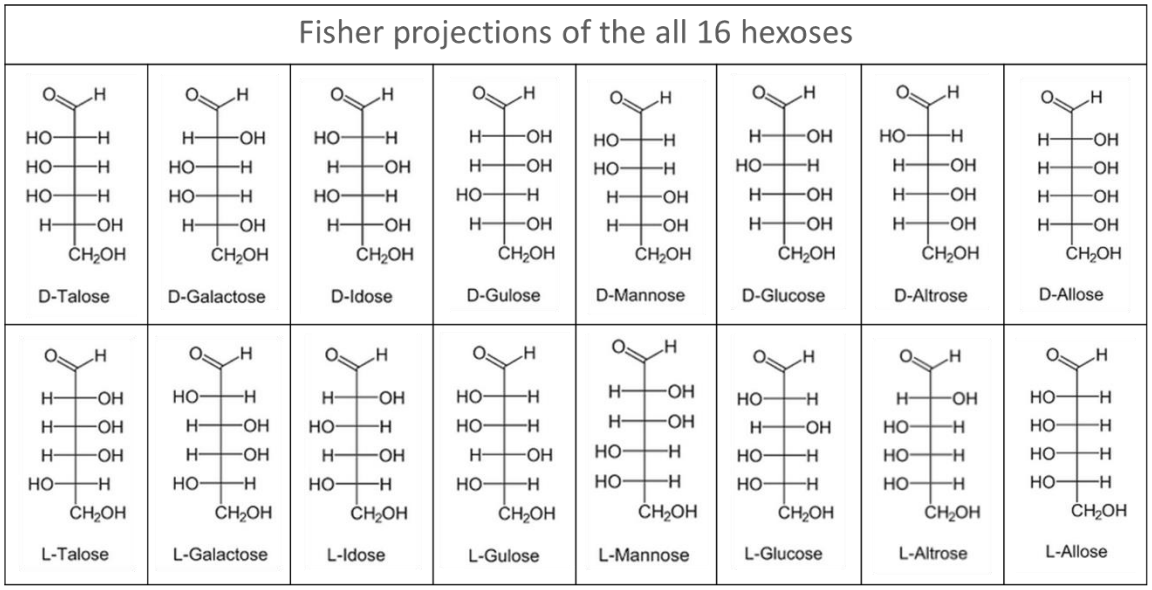
Chapter 1- General Introduction and Objectives

1.1 Introduction to carbohydrates

Initially, carbohydrates were usually associated with cell metabolism context and energy production. The evolution resulting from the sequencing of the human genome (and of other organisms) left lipids and carbohydrates excluded [Varki, *et al*, 2009], with their crucial biological roles on physiological systems to be explained.

Carbohydrates are mono-, di-, oligo- and polysaccharides. Monosaccharides are the simplest carbohydrates. Giving an example the glucose, it is a hexose (6 carbon atoms). However, 4 of their 6 carbons are chiral, thus 16 chemical configurations are possible.

Table 1.1- Illustration of 16 possible structures of hexoses by fisher projections.



For example, only one difference on the C4 atom configuration results in 2 monosaccharides: glucose and galactose (figure 1.1) [Perez, 2014].

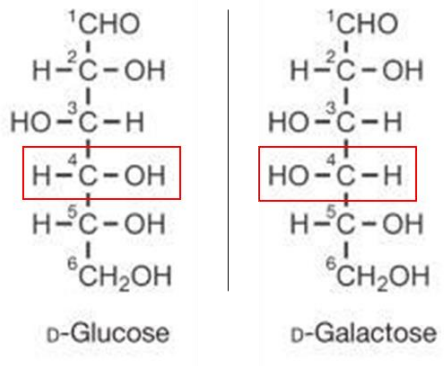


Figure 1.1- Stereochemistry of glucose and galactose. Red rectangle evidences the C4 of each structure which is a mirror image.

Adding to the chirality of the monosaccharides, other modifications can occur. Continuing with the glucose structure, alteration of the 2-hydroxyl group with an acetylated amino group forms N-acetylglucosamine (GlcNAc), or the oxidation of C6 results in a carbohydrate acid, the glucuronic acid (GlcA) [Maureen, *et al*, 2003].

The monosaccharides by itself have a great diversity. The linkage between the monosaccharides increases this diversity, which some of the structures will be described.

Disaccharides are other type of carbohydrates. They are composed of 2 monosaccharides joined by a glycosidic linkage, through a condensation reaction. However, from the same monosaccharides, several disaccharides structures are possible [Varki, *et al*, 2009]. Here we show as examples 3 glucose disaccharides in the figure 1.2.

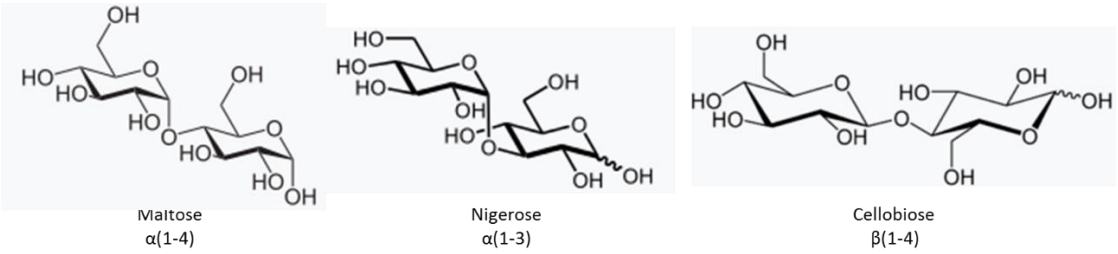


Figure 1.2 - Schematic illustration of structures of 3 glucoses disaccharides.

There are two variables in this context: 1) the glycosil linkage between two carbon atoms 2) the conformer of the anomeric carbon (C1 atom) in α or β.

The formation of different glycosil linkages has an incredibly importance in Biology. Malectin, for example, is able to recognize maltose and nigerose, although has a preference for nigerose due to the conformation of the disaccharide linkage in α -(1-3). On the other hand, malectin has a weaker interaction with cellobiose. Although maltose and cellobiose have the same carbons involved in the glycosylic linkage, the anomeric carbon has a different conformer, thus conferring the structure a different arrangement (further information about the malectin is in the section 1.2) [Schallus, *et al*, 2008].

Oligosaccharides have a number of monosaccharides usually less than 12 and can be covalently linked to macromolecules (glycoconjugate). As glycoconjugates, they have several biological roles such as in cell-cell interactions as well as for host-microbe interaction (figure 1.3), either through symbiotic relationships or pathogenicity [Varki, *et al*, 2009].

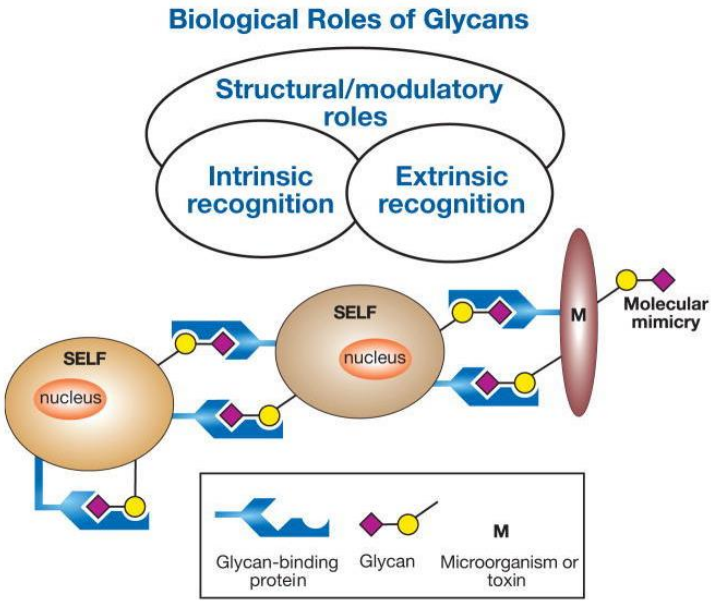


Figure 1.3- Illustration of the intrinsic and extrinsic glycan binding proteins interactions. Image adopted from Essentials of Glycobiology book, chapter 6: Biological Roles of Glycans.

Polysaccharides, in turn, are long chains of monosaccharides units bound together (usually more than 12 monosaccharides) and their structure can vary from linear to highly branched. Due to their large size, they are often insoluble in water [Varki, *et al*, 2009].

The biological roles of polysaccharides are in storage or structural. Glycogen and amylose are examples of storage polymers which are glucoses monosaccharides joined by alpha-linkages (figure 1.4; A and B).

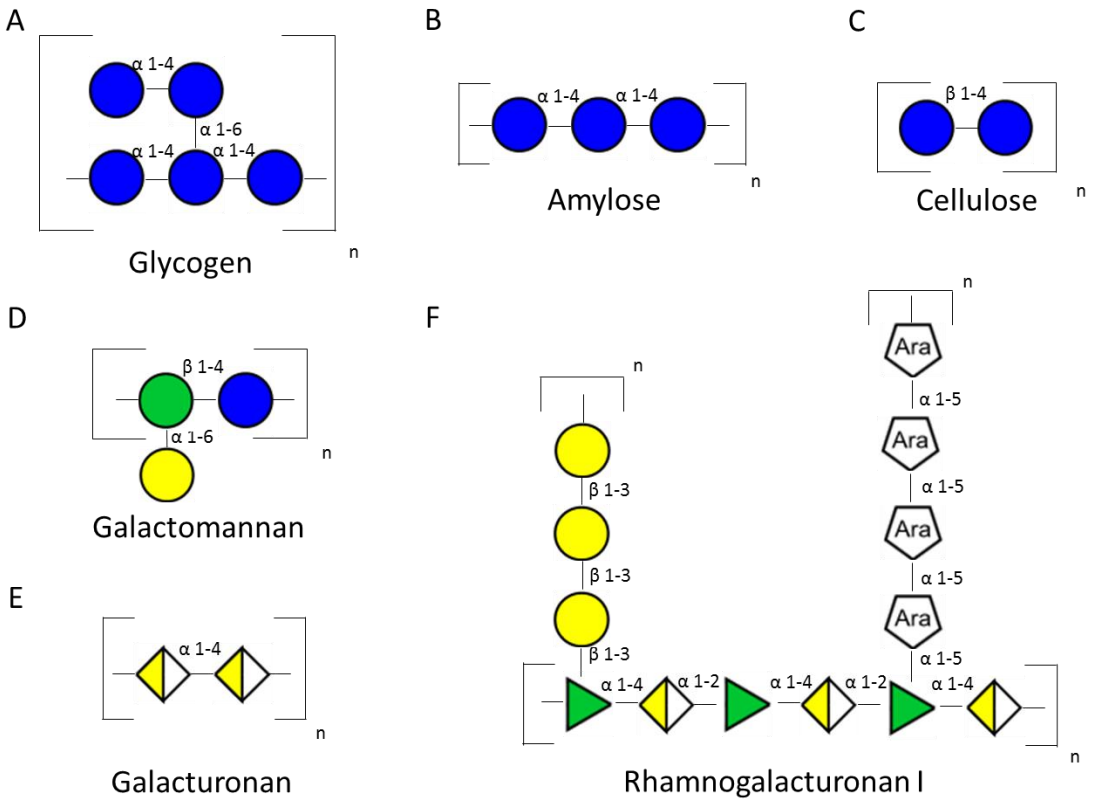


Figure 1.4- Illustration of several carbohydrates in nature, by their symbol nomenclature. A and B- Storage macromolecules of glucose in animals (glycogen) and plants (amylose). C- Cellulose, a cell wall component of plants. D-Galactomannan, a hemicellulose, other cell wall component of plants. E and F- Pectins, from the simple structure (galacturonan) to high complex structures (rhamnogalacturonan I).

Cellulose is another polysaccharide that exists in plant cell walls, composed of glucose monosaccharides joined together by β-linkages (figure 1.4; C) [Bledzki, *et al*, 1999]. Human enzymes lack the specificity for β-linkage cleavage, thus humans are unable to degrade this polysaccharide. More than celluloses, the plants cell walls may have hemicelluloses and pectins (figure 1.4; D, E and F), whose macromolecules are more complex. Hemicelluloses besides glucose includes xylose, mannose, galactose, rhamnose and arabinose [Scheller, *et al*, 2010]. Pectins, in turn, have in their backbone galacturonic acid. However, pectins can have several types of monomers decorating their backbone, hence their incredible diversity. Pectin can be simple structure such as galacturonans, which are linear chains of galacturonic acid linked by α-(1-4) linkage, or complex polysaccharides such as rhamnogalacturonan I (RG I). The RG I besides the galacturonic acid in the backbone, it has also α-(1-2) rhamnose. In addition, it has several branches that contain arabinan and galactan. Furthermore, there is another extremely complex pectin, rhamnogalacturonan II. It comprises a homogalacturonan backbone with side chains comprising 12 different monosaccharides linked by 20 linkages type. Although more complex than RG I, its size is smaller [Gorshkova, *et al*, 2010].

1.2 Malectin

Malectin was first found and described in *Xenopus laevis* pancreas, but it was soon revealed to be present in other tissues and organisms [Schallus, *et al*, 2008].

Malectin is a membrane-anchored reticulum endoplasmic protein involved in the N-glycosylation pathway (a process in which the nascent glycoproteins have the assisted folding). It has a β-sandwich fold (figure 1.5) capable of binding to nigerose, maltose and di-glucosylated oligosaccharide, with nigerose as preferred ligand [Schallus, *et al*, 2008].

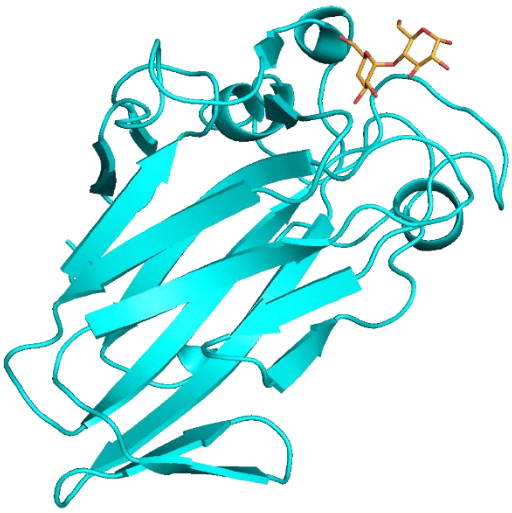


Figure 1.5 - Tertiary structure of malectin with beta-sandwich fold, solved by X-ray crystallography technique. Pinheiro, *et al*, unpublished.

The important residues that have been observed for making carbohydrate interactions are: Ser80, Glu102, Lys138, Asp201 and Asn202 through direct hydrogen bonds; Tyr82, Glu129, Val130 and Glln137 through water mediated hydrogens bonds; Tyr104 and Tyr131 through π-CH interactions [Pinheiro, *et al*, unpublished].

For a better understanding of this protein role, a brief introduction will be made for the N-glycosylation pathway. The nascent glycoproteins are unfolded and are transported into the ER lumen. There, an oligosaccharide with 3 terminal glucoses is transferred to the protein with Asn-X-Thr/Ser sequence consensus [Lehle, *et al*, 2006]. In short, 2 glucoses are successively hydrolysed by 2 α-glucosidases, and 2 chaperones proteins bind to mono-glucosylated proteins, initiating the assisted-folding. After the process is completed, the glycoproteins are released from the chaperone proteins, the α-glucosidase hydrolyses the third glucose residue and the glycoproteins are then transported to the Golgi apparatus (figure 1.6) [Deprez, *et al*, 2005].

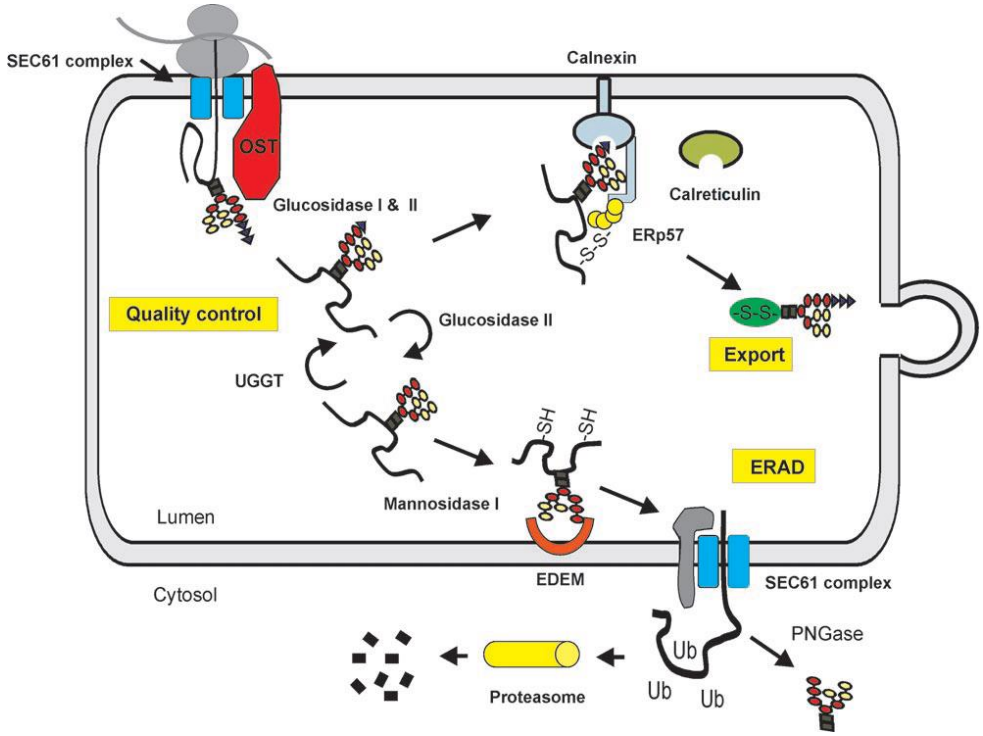


Figure 1.6- Illustration of the mechanism of nascent protein assisted-folding. Glucosidases I and II hydrolyses two glucoses residues of oligosaccharide, so this glycoprotein is recognized by chaperones proteins Calnexin and Calreticulin. Image adopted from *Protein Glycosylation, Conserved from Yeast to Man: A Model Organism Help Elucidate Congenital human diseases* review.

Malectin participates in this process before the hydrolyse of the second glucose. Here, the malectin is suspected to function as a quality control. In a study, where the cells were transfected with the DNA encoding hemagglutinin to enhance the stress of the ER, it showed that the expression of the malectin increased [Galli, *et al*, 2011]. Since there is a delay in the hydrolysis of the second glucose, the malectin can protect these glycoproteins to prevent their degradation [Schallus, *et al*, 2010]. Furthermore, the same study has shown the Malectin doesn´t affect the kinetic of the chaperone lectins but might identify the misfolded glycoproteins for their degradation, thus preventing the secretion of misfolded proteins to the Golgi and avoiding the formation of aberrant proteins [Galli, *et al*, 2011].

Interestingly, the malectin sequence shares similarity with some sequences of carbohydrates binding modules (CBMs) [Schallus, *et al*, 2008].

1.3 Carbohydrates-binding-modules (CBM):

Carbohydrate-Binding Modules (CBMs) are proteins with the ability to bind to glycans. CBMs were originally classified as cellulose-binding, but the term soon evolved with the discovery of new CBMs that had other specificities [Shoseyov, *et al*, 2006]. Although they are present in all kingdoms of life, they predominate in prokaryotes.

CBMs are defined as non-catalytic modules that are appended to carbohydrate-active enzymes [Shoseyov, *et al*, 2006]. CBMs have a range of 30 to 200 amino acids, are in their majority composed of beta-sheets and can be found as single or multiple modules in one protein. CBMs are usually present in N- or C-terminal, but have already been observed in the middle of the protein sequence. The recognition of glycans is done with aromatic amino acids chains by π-CH interactions or with polar amino acids that are capable to form hydrogen bonds (figure 1.7). Some of these CBM modules are calcium ion-dependent, to stabilize the interactions (figure 1.7) [Boraston, *et al*, 2004].

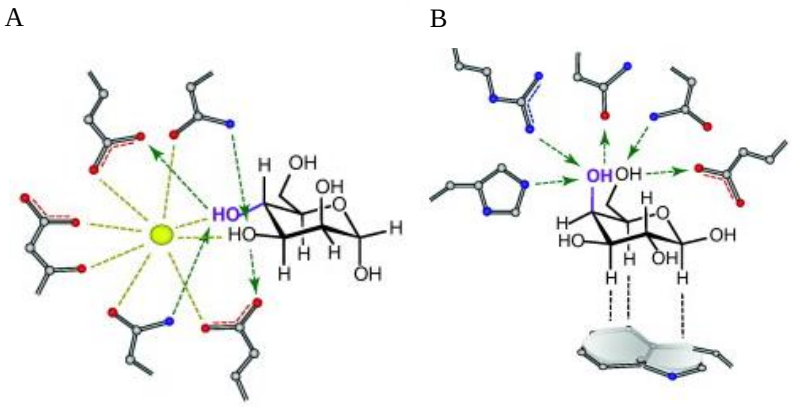


Figure 1.7- Two types of interactions with a carbohydrate. A-Hydrogen bonds between polar amino acids chain to hydroxyl group with the cordination of the calcium. B-Hydrogen bonds between polar amino acids chain to hydroxyl group and aromatic amino acid chain reoriented for π/C-H interaction. Arrows represents hydrogen bonds. Yellow circle represents calcium cation. Image adopted from *Lectin structure to functional glycomics: principle of the sugar code* article.

The affinity of the CBMs for carbohydrates is relatively low, generally around 10^5-10^4 M^{-1} [Schallus, *et al*, 2010]. Studies have shown that the addition of a CBM in a protein increased the catalytic activity while the removal of the CBM decreased the efficiency [Shoseyov, *et al*, 2006].

However, CBM modules are not always appended to catalytic modules, but can be individualized in one single protein. Occasionally, a CBM module can have multiple binding-sites [Boraston, *et al*, 2004].

1.3.1 Classification of CBMs:

Nowadays, there are 81 CBM families in CAZy database based on their sequence similarity, with 7 possible folds: β-sandwich, β-trefoil, Cystein knot, Unique, OB fold, Hevein fold or mixture of Uique and Hevein fold. The β-sandwich is so far, the major structure of CBMs. It comprises two β-sheets, each with 3 to 6 antiparallel β-strands. In addition, the CBMs are divided into three different types accordingly to their binding typology (figure 1.8): 1) type A, glycan surface-binding; 2) type B, glycan chain binding; 3) type C, small sugar binding [Boraston, *et al*, 2004].

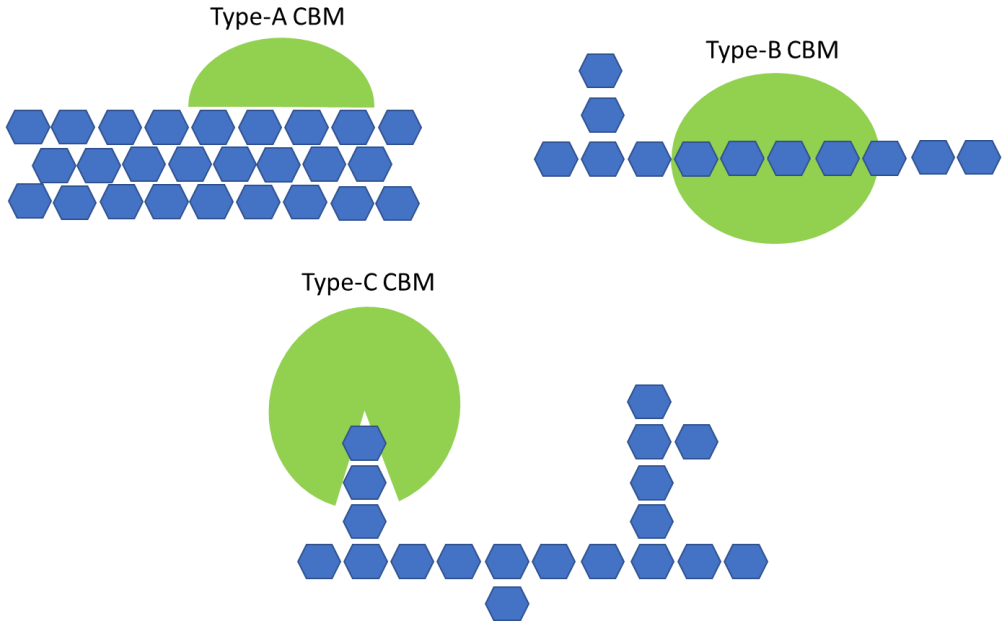


Figure 1.8- Schematic illustration of 3 types of CBMs with different binding to polysaccharides.

The type-A CBMs have the most distinct binding-interaction, with a planar binding site that binds to the surfaces of cellulose or chitin. The type-B CBM have groves that can accommodate various glycan units, while the type-C CBMs have short pockets for recognition of mono-, di- or tri- saccharides [Boraston, *et al*, 2004]. However, in type-C CBMs, the solved structure doesn´t imply knowledge of the function, since members of the same CBM family may have different specificities [Taylor, *et al*, 2014].

In addition, the CBMs modules can have 4 different roles: 1) targeting effect; 2) proximity effect; 3) disruptive effect; 4) adhesion. In the targeting effect, the CBM target the enzyme to distinct zones within a larger glycan, either at the terminals or internal parts of polysaccharides chains. In the proximity effect, the CBMs directs the substrate to the enzyme to increase the efficiency of polysaccharide degradation. In the disruptive effect, the CBMs disrupt the surface of packed polysaccharides (fibres or granules), exposing the substrate more easily to the catalytic module. In the adhesion, the CBMs adhere onto the surface of the bacterial cell wall glycans while the enzyme has an activity on neighbour glycans [Shoseyov, *et al*, 2006].

1.4 Human gut microbiome:

Recently, several CBMs that share sequence homology with malectin were classified by CAZy as the novel CBM57 family. In addition, members of CBM6 and CBM35 families, both type-C CBMs (although family 6 CBMs can also be from type-B) also share sequence similarity with members of the CBM family 57. These CBMs exists in the genome of bacteria, some of them in our gut.

Human microbiome is composed of several bacteria microorganisms that live in our bodies, a superior number of 10 times more than our cells [Lynch, *et al*, 2016]. These microorganisms have several impacts in the physiologic functions for our health and disease, since they protect us against pathogens and educate our immune system. The human gut has the most diverse population, with Bacteroidetes and Firmicutes as dominant phyla [Shreiner, *et al*, 2015]. Here, we will focus on Bacteroidetes phylum.

1.4.1 The role of the Bacteroidetes in the human gut

Our organism can easily absorb carbohydrate monomers and cleave some disaccharides, but several polysaccharides, such as plant cell walls components, are resistant to our digestive enzymes (example of cellulose given in the section 1.1). Members of the Bacteroidetes phylum have the necessary tools for the degradation mechanism of these polysaccharides, thus complementing the eukaryotic genome [Thomas, *et al*, 2011].

B. thetaiotaomicron and *B. ovatus* are most known for adopting the Polysaccharide Utilization Loci (PULs) system as a strategy for glycans degradation. *B.thetaotaomicron* has several PULs that *B.ovatus* doesn´t have, but on the other hand, *B.ovatus* has unique PULs for hemicelluloses degradation [Martens, *et al*, 2011]. Since these 2 bacteria are in distinct zones of the gut, this characteristic may due to the adaptation of nutrient availability.

Here, we show one PUL organization of *B.thetaiotaomicron* in figure 1.9, which includes some CBM modules used in our studies.

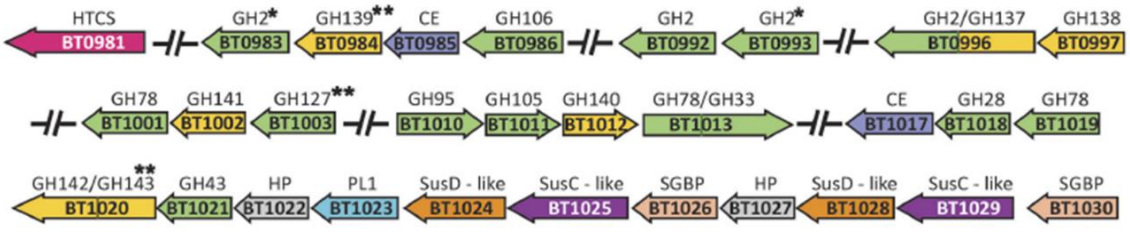


Figure 1.9- Organization of genes that encodes proteins (and their orientation) for the degradation of RG II. Image adopted from *Complex pectin metabolism by gut bacteria reveals novel catalytic functions* article.

This PUL is specific for RG II degradation, which encodes a starch-utilization system-like (Sus-like), a polysaccharide lyase, carbohydrate esterases and several glycoside hydrolases [Ndeh, *et al*, 2017]. The Sus-like system is composed of at least one pair of outer membrane proteins homologous to SusC and SusD, crucial for the transport and degradation of polysaccharides [Shipman, *et al*, 2000]. In periplasm, the polysaccharide lyase cleaves the backbone of the RG II, proceeded by the action of carbohydrates esterases and glycoside hydrolases, which is coordinated [Ndeh, *et al*, 2017]. Carbohydrates esterases hydrolyse the esters group into acid and alcohol groups [Nakamura, *et al*, 2017], while the glycoside hydrolases cleave the specific glycosidic linkage [Naumoff, *et al*, 2011].

The CBM modules from this PUL, and used in our study, are appended to the Bt0996 protein that has 2 distinct catalytic modules: 1) a glycoside hydrolase family 2 with β-D-glucuronidase activity; 2) a novel glycoside hydrolase family 137 with β-L-arabinofuranosidase activity. The CBM modules, as mentioned in the section 1.3.1, can have different roles and different specificities.

1.5 Objectives:

Decoding the specificities of the novel members of the CBM57 family and homologues (CBM6 and CBM35 families) is important for a better understanding of the evolutionary relationship between the malectin, a well conserved lectin in animal domain, with these modules identified in prokaryotes.

Moreover, these modules are present in two bacteria of the human gut, appended to glycoside hydrolases. Assigning their specificities is a crucial step to understand efficient degradation of complex polysaccharides such as Rhamnogalacturonan II. However, the identification of the binding specificity to these modules and the reveal of their roles is challenging.

Here we addressed the problem using first bioinformatic tools to observe the level of amino acid conservation, in particular of the putative binding residues and second to select the CBM modules, for combining glycan microarray with X-ray crystallography to study glycan-CBM interactions.

This Thesis has the aim to elucidate glycan specificities of identified CBMs in the genomes of two human gut bacteria: *Bacteroide ovatus* and *Bacteroide thetaiotaomicron*. The main goals of this thesis are summarized in the following topics:

- To investigate the amino acid sequences using bioinformatic tools and compare with the human malectin sequence to predict carbohydrate-interacting residues and select some CBMs of the human gut bacteria for bio-characterization studies. This is detailed in Chapter 2.

- Re-cloning of the recombinant proteins and use of various expression tests to discover the best condition for a larger-scale production of CBMs, followed by a purification process. This is detailed in Chapter 3.

- To perform two types of glycan microarrays with the successfully produced CBM modules. First, to identify possible novel glycans and second, to determinate the epitope recognition of the CBM modules. This is detailed in Chapter 4

- To perform crystallization assays to explore the 3D structure and compare with the malectin structure. This is detailed in the Chapter 5

The results obtained in this thesis were presented as poster communications in 2 Meetings and are included in Index.

2017-May, Lourenço, F; Leote, C; Correia, VG; Brás, JL; Fontes, CMGA; Romão; MJ; Carvalho, AL; Palma, AS; Pinheiro, BA. Assignment of biological roles for CBM57 modules domains in human gut bacteria. 6th Meeting of Portuguese Synchrotron Radiation Users. LNEG, Lisbon.

2017-September, Lourenço F, Correia VG, Leote C, Brás JL, Fontes CMGA, Romão MJ, Carvalho AL, Palma AS, Pinheiro BA. Decoding Type-C CBM modules specificities in human gut bacteria. 12th Meeting of the Group of carbohydrates. University of Aveiro, Aveiro.

Chapter 2- Phylogenetic studies

2.1 Introduction:

The accumulation of mutations in genes encoding proteins is an evolution mechanism. Usually the mutations are functionally and structurally silent, not having influence in the final protein. In some amino acids, the change may cause function loss, being the mutation called deleterious and eliminated by natural selection. On the other hand, the residue change may be neutral (similar residue) or the protein activity increased, and the mutation accumulated in the protein sequence, relating the evolution history [Soskine, *et al*, 2010].

Aligning molecular sequences (protein or DNA) of homologous proteins from all kingdoms of life helps to understand their evolution history, further helping the identification of conservation and variation sites [Yang, *et al*, 2012].

For several proteins (modules), the structure and, specially, the function role remains unknown. Phylogenetic studies help the prediction of structurally homologous proteins and the alignment of the molecular sequences, the identification of conserved and varied sites, help the prediction of the conservation (or not) of the protein function and specificity. Phylogenetic studies use bioinformatic tools (molecular sequence alignments and tree diagrams) to understand the gene evolution on various living organisms.

The tree diagram (figure 2.1) is composed of branches, and their sizes describe the molecular sequence divergences. The tips correspond to the molecular sequence of living organisms (taxa) while the nodes represent their respective ancestor and the root node the common ancestor [Yang, *et al*, 2012].

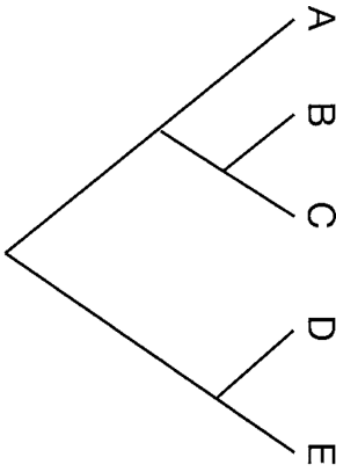


Figure 2.1- An example of diagram tree. A-E represent taxa, internal nodes represent the ancestor and the root represents the common ancestor for all the taxa. Image adopted from *Essential Bioinformatics* book, chapter 10.

The tree diagram itself is the topology, the relationship of each taxon displayed in the phylogenetic tree (which taxon descended first and the node that belongs). It is possible in the final to have more than one topology, because various solutions are plausible to describe one evolution history [Yang, *et al*, 2012].

Considering that the protein evolution is divergent, if the ancestor descends to two taxa, then the tree diagram must bifurcate. However, if the tree diagram is multifurcating (the previously ancestor descends to more than two taxa), it is difficult to predict which of the two taxon descended first or if three taxa were instantly descended from the ancestor, instead of two taxa only. This process is known as radiation [Xiong, *et al*, 2006]. In addition, the generated tree can be rooted or unrooted. In the case of a rooted phylogenetic tree, it is assumed that all taxa have a common ancestor (and a root node is showed up) [Yang, *et al*, 2012].

In case the number of molecular sequences is large, more topologies are possible and the branch size estimation isn´t always the correct one, being the phylogenetic study a complex task, demanding more computation [Yang, *et al*, 2012].

For phylogenetic tree construction, first the type of molecular sequences must be chosen (nucleotides or amino acids sequences) and then retrieved from the databases.

When related organisms are studied, the mutations in the molecular sequence are relatively low and for a better evolution history, DNA sequences should be used (the gene evolves faster than the protein) [Xiong, *et al*, 2006]. On the other hand, if the organisms in study are divergent, the amino acids sequences should be used, otherwise it will be too complex to understand the evolution [Yang, *et al*, 2012].

The molecular sequences can be retrieved from diverse databases, for example, CAZy or SwissProt. The function prediction of each protein can be achieved by using the InterPro Scan program, which identifies the protein domains. This tool analyses each retrieved sequence individually and compares to sequences from the databases called signatures. After the analyse, the program presents the predicted domains [Jones, et al, 2014]**.** In case we want to clone a DNA with right nucleotides, other programs must be simultaneously used to see the agreement, or otherwise, the gene cloned could be codifying a truncated protein.

Obtained the sequences, alignments programs are used to analyse and compare them. When constructing a phylogenetic tree, it is important to use several state-of-the-art alignment programs such as ClustalOmega, T-Coffee or Muscle. Wrong alignments usually lead to a misinterpretation of an adaptive evolution process and, consequently, to an untruthful phylogenetic tree. The alignment programs align amino acid sequences by homology and add gaps in amino acids sequences when necessary, making sure that the amino acids sites aligned are identical (by chain propriety) as much as possible [Xiong, *et al*, 2006].

Giving ClustalOmega as example of a multiple alignment program function, the distance matrix is calculated using all taxa sequences and calculating a guide tree. The closest sequences pairs related in the guide tree are aligned. Each sequence pair aligned is considered as a consensus sequence and alignments of every consensus sequence are performed, generating a new consensus sequence. The sequences unpaired are aligned to the consensus sequence and the alignment performed by ClustalOmega is completed (figure 2.2) [Xiong, *et al*, 2006].

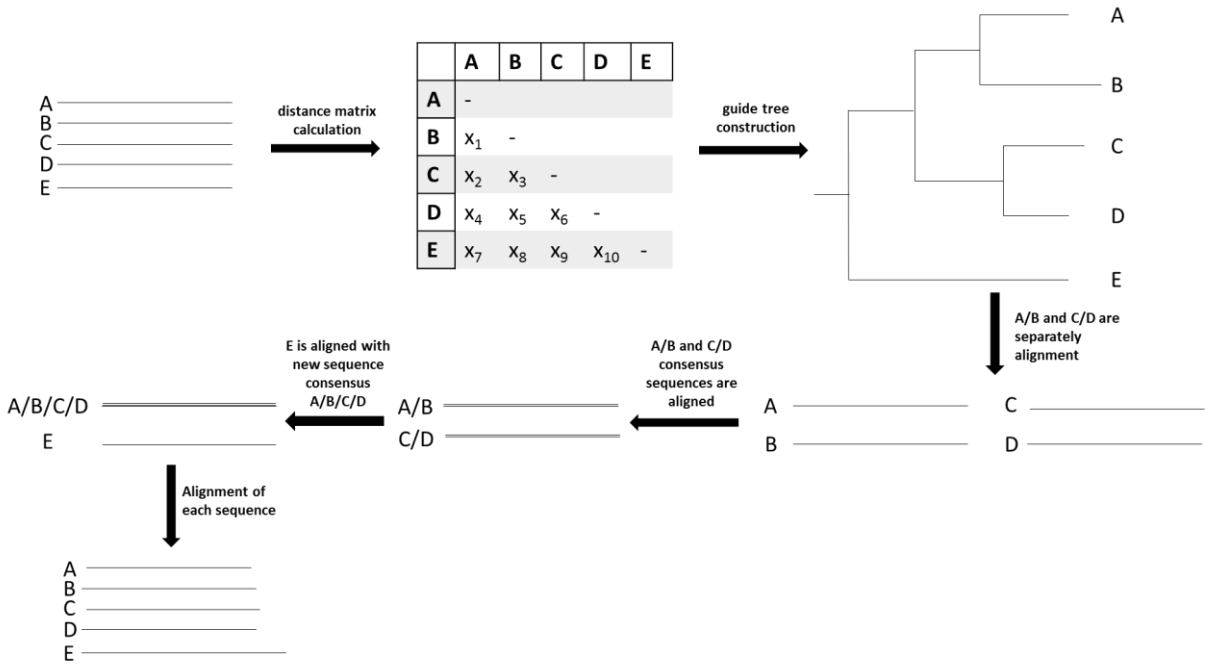


Figure 2.2- Illustration the function of ClustalOmega (a multiple alignment program).

The addition of gaps has a penalty score. Penalty scores allow the insertion of gaps more easily in variation regions, for example loops, than in the conserved regions. Also, the gap insertion between hydrophobic residues increases that penalty score [Xiong, *et al*, 2006].

There are two categories in the construction of phylogenetic tree: one based on distances, which calculates the distance matrix and one based on discrete characters (analyses each character site) [Yang, *et al*, 2012].

2.1.1 Distance methods:

Distance is defined as a count of substitutions present in two sequences aligned. Distance methods analyses the evolutionary distance between the sequences using the scores in the alignment to construct the distance matrix and derive the phylogenetic tree. The most common distance methods are UPGMA and Neighbor-Joining, that are detailed below [Xiong, *et al*, 2006].

2.1.1.1 Unweighted Pair Group Method Using Arithmetic Average (UPGMA):

UPGMA is the easiest distance method. As mentioned above, the distance matrix is calculated and the two sequences with the smallest value are joined together and a node between them is added. The pair sequence is treated as a consensus sequence and the distance matrix is recalculated. Again, another molecular sequence is clustered and the cycle is repeated until the last molecular sequence taxon is joined in the tree, creating a rooted tree [Xiong, *et al*, 2006].

UPGMA assumes that the molecular sequences evolve at constant rate, proportional to the accumulation of mutations (known as Molecular Clock technique). However, mutations usually don't occur at a constant rate, which can lead to a UPGMA phylogenetic tree creation that is not entirely true. Despite this limitation, it is still a very fast method for constructing a phylogenetic tree [Xiong, *et al*, 2006].

2.1.1.2 Neighbor-Joining method

Neighbor-Joining method is another option for using a distance method for the construction of phylogenetic trees. Opposite to UPGMA, it uses an evolutionary rate correction formula [Xiong, *et al*, 2006].

The Neighbour-Joining tree is based on the minimum evolution principle, where the branch length must be the lowest as possible. In the beginning, the tree has a star form with all branch lengths at the same value and the sum of their lengths is then calculated. Then, a pair of taxa is clustered together into a neighbour (node) with the sum of all branches length being recalculated. The pair of taxa with smaller sum value is joint in the node. This process is repeated until the final tree is generated (figure 2.3) [Xiong, *et al*, 2006].

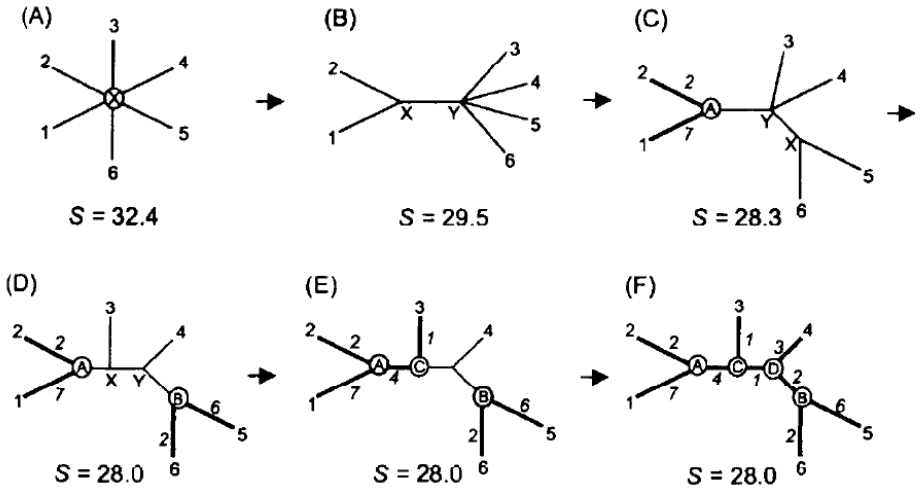


Figure 2.3- Neighbour-Joining tree phylogenetic construction process. The final tree has lower sum of branch length than the beginner tree (star tree). Image adapted from *Essential Bioinformatics* book, chapter 10.

The final constructed tree must have the lower sum of branch lengths. The branch lengths between nodes is representative of the amino acids differences between the sequences [Xiong, *et al*, 2006].

The main limitation is that only a single topology is analysed. When trying to understand the evolution stages from (highly) divergent species, the final tree may not be completely correct. However, its advantage is the same as in UPGMA, since it has a fast computation time [Yang, *et al*, 2012].

2.1.2 Discrete character methods:

Discrete character methods analyse directly the character sites (nucleotide or amino acid site) in the alignment sequences and don´t involve distance matrix calculations. Maximum-Parsimony (MP) is one of the most known method based on discrete character [Yang, *et al*, 2012].

Maximum-Parsimony analysis molecular sequences sites with large divergence due to the information given about the evolution (mutations in the site). The non-informative sites are scrapped. The amount of changes at each informative site is summed and every possible phylogenetic tree is calculated [Yang, *et al*, 2012].

The discrete character method, opposite to the distance methods, gives the information about the molecular sequence from the common ancestor, which it is already extinct and impossible to retrieve [Yang, *et al*, 2012].

Every possible phylogenetic tree has the sequence of his own molecular common ancestor, with the most plausible solution being the phylogenetic tree that requires the minimum number of substitutions in the molecular sequence. This is done by analysing each informative site. For example, if the molecular sequence used is DNA and the taxa in a specific informative site have guanine, thymine or adenine, the nucleotide present in the ancestor common is the one requiring less substitutions. In a site, if six taxa are analysed and three have thymine, two have adenine and one has guanine, for the minimum substitutions the common ancestor sequence site must be thymine (figure 2.4) [Yang, *et al*, 2012].

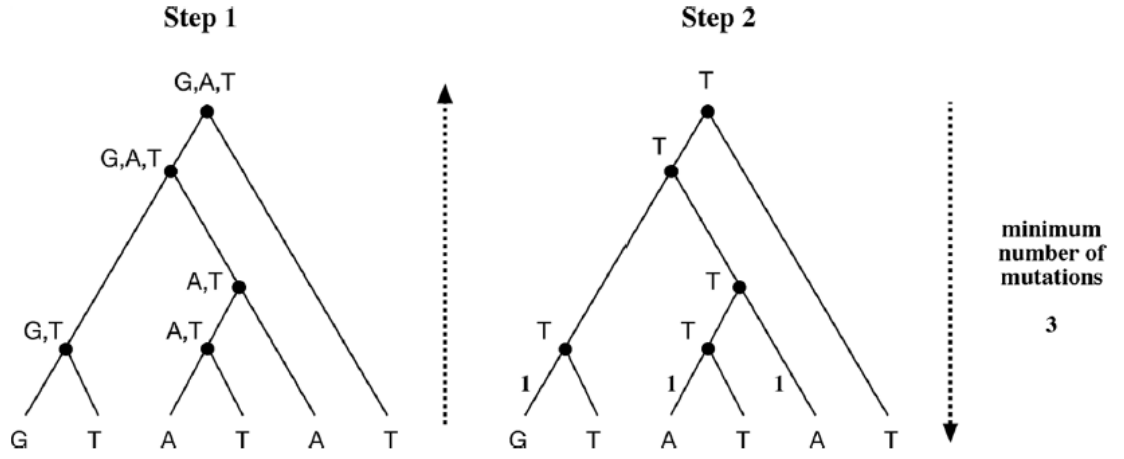


Figure 2.4- Illustration of Maximum Parsimony mechanism, predicting the ancestor characters in two steps. The first step is to count all possible characters at each node. The second step is choosing the character in common ancestor that involve minimum number of substitutions. In this case, the common ancestor character must be T so the total number of substitutions is minimum. Choosing any other character would lead to an increase of substitutions. Image adopted from *Essential Bioinformatics* book, chapter 11.

The obvious advantage is the molecular sequence information from the ancestor. Moreover, this method produces more reliable phylogenetic trees, in comparison of distance methods. The biggest limitation is it requires a large amount of computation time, that exponentially increases when the amount of molecular sequences to be analysed is extended (there are more possible topologies). Here, the molecular sequences amount shall not exceed the number of ten. Other counterpart is it assumes all substitutions mutation types are at the same frequency, when it is known that some substitutions happen more frequently than others [Yang, *et al*, 2012].

The following table describes in general the pros and cons of each method referred above:

Table 2.1- List of pros and cons from all the three methods described above.

Name	Based method	Pros	Cons
UPGMA	Distance	-Fast computation -Easiest of all methods	-Assume the evolution rate is constant (Molecular Clock)
Neighbour Joining	Distance	-Fast computation -Do not assume the evolution rate is constant	-Constructs only a possible tree
Maximum Parsimony	Discrete character	-More accurate phylogenetic trees compared to the methods based on distance	-Computation time demanding -Cannot be applied to high divergence sequences

2.1.3 Phylogenetic tree validation:

After the phylogenetic tree construction, there are statistic criteria to evaluate the confidence of the made tree, such as bootstrap, jackknifing and others [Yang, *et al*, 2012].

Bootstrap has the principle of causing perturbations in the alignments, which can be done by the random replacement of sites. From here, there is a re-alignment and reconstruction of the phylogenetic tree (using the previously used method). This test is used 100 to 1000 times and the phylogenetic trees are clustered into a consensus tree. Bootstrap relies in the assumption that if the phylogenetic tree is strongly robust, the bootstrap consensus tree will be identical to the original tree and the evolution history is statistically correct. Moreover, the statically values added above each node in the phylogenetic tree represents the confidence of two taxa descended from the same ancestor. Values equal or superior of 70% are considered with confidence [Yang, *et al*, 2012].

Jackknifing is another validation method. This method relies in randomly scraping-out half of the sites and re-aligning and reconstructing the phylogenetic tree. However, using this approach has the counterpart that the sequence alignments are shorter and thus the original tree may no longer be replicated [Yang, *et al*, 2012].

Molecular Evolutionary Genetic Analysis (MEGA) is one of the programs for phylogenetic tree construction. It has the option of using DNA or protein alignments, constructing the phylogenetic tree using one of the three methods mentioned. The validation is then done by performing 100 to 1000 bootstrap tests [Kumar, *et al*, 2016]. If necessary, it is possible to edit the obtained alignment for a better phylogenetic tree construction.

2.2 Materials and Methods

To evaluate the divergent evolution of CBM57, 315 protein sequences from bacteria and eukaryote were retrieved from the carbohydrate active enzymes (CAZy) database to date (06/09/2016). The domain architecture determination of each protein was performed using the InterProScan program [Jones, *et al*, 2014].

The alignments were done using the program Clustal Omega [Sievers, *et al*, 2011]. To facilitate interpretation, the alignment of three groups of protein sequences were made: bacterial, plants and eukaryotes (non-plants). The alignments were then analysed using the tertiary structure of the human malectin domain. For checking the conserved regions, the software ESPript 3.0 [Robert, *et al*, 2014] was used. Since the binding site is present in the loops, these were carefully analysed to detect amino acid changes and predict possible specificity differences. In addition, phylogenetic trees were done for checking the evolution of malectin through-out all kingdoms of life.

Since the alignment of protein sequences gives a higher signal-to-noise in alignments and phylogenetic analysis, they were used to study the evolution of malectin in this widely divergent group of organisms (bacteria and eukaryotes). Also, we had the advantage of having the tertiary structure of two of the proteins to be used in the study, which served as guide lines for the alignment. For improvement, highly divergent sequences were removed from the alignment.

The phylogenetic trees were done using the software MEGA7.0 [Kumar, *et al* 2016]. The Neighbor-Joining tree building method was used for phylogenetic tree construction. This method allows the analysis of several divergent protein sequences at the same time, using small computation times when compared to other tree construction methods, but still giving a satisfactory final tree.

The variation of amino acid residues in the sequence was estimated by the p-distance model, the simplest distance measure, which the value is a division of the number of amino acids variation by the total number of amino acids compared [Nei, *et al*, 2006]. Gaps presented in amino acids sequences were treated using the pairwise-deletion option (the total deletion option loses some information about gene evolution) and the validation of phylogenetic tree was done by performing 1000 bootstrap replications.

2.3 Results and Discussion

2.3.1 Analysis of 315 peptides sequences in different species:

Understanding the evolution of malectin in the different species from all domains of life, may reveal important to predict novel specificities for CBM57 family members and the adaptive nature of this carbohydrate-protein interaction. These malectin-like modules are present in most of the kingdoms of life. They are found in Archaea (1), prokaryotes (176) and eukaryotes (140).

In plants, the malectin-like module is associated with kinases, being probably involved in signalling and regulatory processes. In other eukaryotes, the malectin-like module is an entire protein, probably involved in the N-glycosylation pathway. Interestingly, the malectin-like module is absent in fungi. One hypothesis is that during the evolution history the fungi kingdom, their members lost the gene coding for malectin.

In prokaryote kingdom, the malectin-like modules have a widespread distribution. These modules are part of proteins present both in Gram-positive and Gram-negative bacteria and in several ecologic niches (water, gut soil). Analysing the architecture of bacterial proteins with the CBM57 module, most have the sequence homology to the membrane transporter Tonb dependent receptor like-module or a catalytic module in their structure (table 1).

These modules (presented in table 2.2) may give a hint of possible specificities for the CBM57 module to which they are appended to.

Table 2.2- Principal domains and function associated with CBM57 family in Bacteria and distribution in life.

Domain	Function	Number of proteins present	Number of organisms present	Type of organisms present
Glycoside hydrolase (GH2)	Hydrolyses the glycosylic bond between carbohydrates	49	38	Gram - Gram+
Pectin Lyases	Degradation of pectin	15	6	Gram -
Peptidase (S8/S52)	Serine Protease; cleaves serine of N- or C-terminal, depending the protein family	10	10	Gram+ Gram-
Quinoprotein alcohol dehydrogenase	Elimination of methanol or ethanol	11	6	Gram+ Gram-

Galactose oxidase/kelch beta propeller	Oxidation of the hydroxyl group at C6 position in galactose	12	10	Gram- Gram+
Polycystic kidney disease (PKD)	Unknown; predicted to interact with others proteins and sugars	20	14	Gram-
TolB-like	Associated for translocation of group A and E colicins that penetrate and kill cells	27	20	Gram- Gram+
PapDlike	Periplasm chaperone that mediates the attachment of bacteria	9	8	Gram+ Gram-
Concanavalin A	Activates proliferation of cells	9	9	Gram -

Most of the modules attached to the CBM57 modules had a function prediction. However, some exceptions were observed. In 8 species of soil bacteria it was observed that the malectin-like modules are individualized, or the InterProScan program couldn´t simply predict the associated catalytic module.

While most of the proteins had only one malectin-like module, some proteins had two or more. They were found in 27 species, majority in soil or water, appended to several putative catalytic modules. More than one malectin-like module may be for increasing the affinity for a specific carbohydrate ligand or each malectin-like module may be for interacting with different epitopes from the same glycan molecule or from different glycan molecules. For this reason, a new set of alignments have been performed to see if there is conservation in the putative interacting residues present in the loop-region of these malectin-like modules.

2.3.2 Evolution of the sequences of the malectin-like modules:

Alignments of the malectin-like modules were performed using selected criteria groups: eukaryotes non-plants, plants and prokaryotes.

In prokaryotes, due to the large number of amino acids sequences that had to be analysed, the strategy adopted was first to construct the phylogenetic tree (using the alignment with all the prokaryotic malectin-like sequences). Since there were many divergent sequences, we used this phylogenetic tree not to describe the evolution pathway (as it seen in the index figure 6, majority of the values on the nodes are below 70%), but rather to see which malectin-like modules clustered together. Modules associated to similar catalytic module were arranged near to each other in the phylogenetic tree, creating sub-divisions. For this reason, new alignments were made using sequences from these sub-divisions.

2.3.2.1 Eukaryotic (excluding plants) malectin-like evolution:

Malectin-like modules from Eukaryotes are highly conserved, having minor differences in their amino acid sequences (figure 2.5).

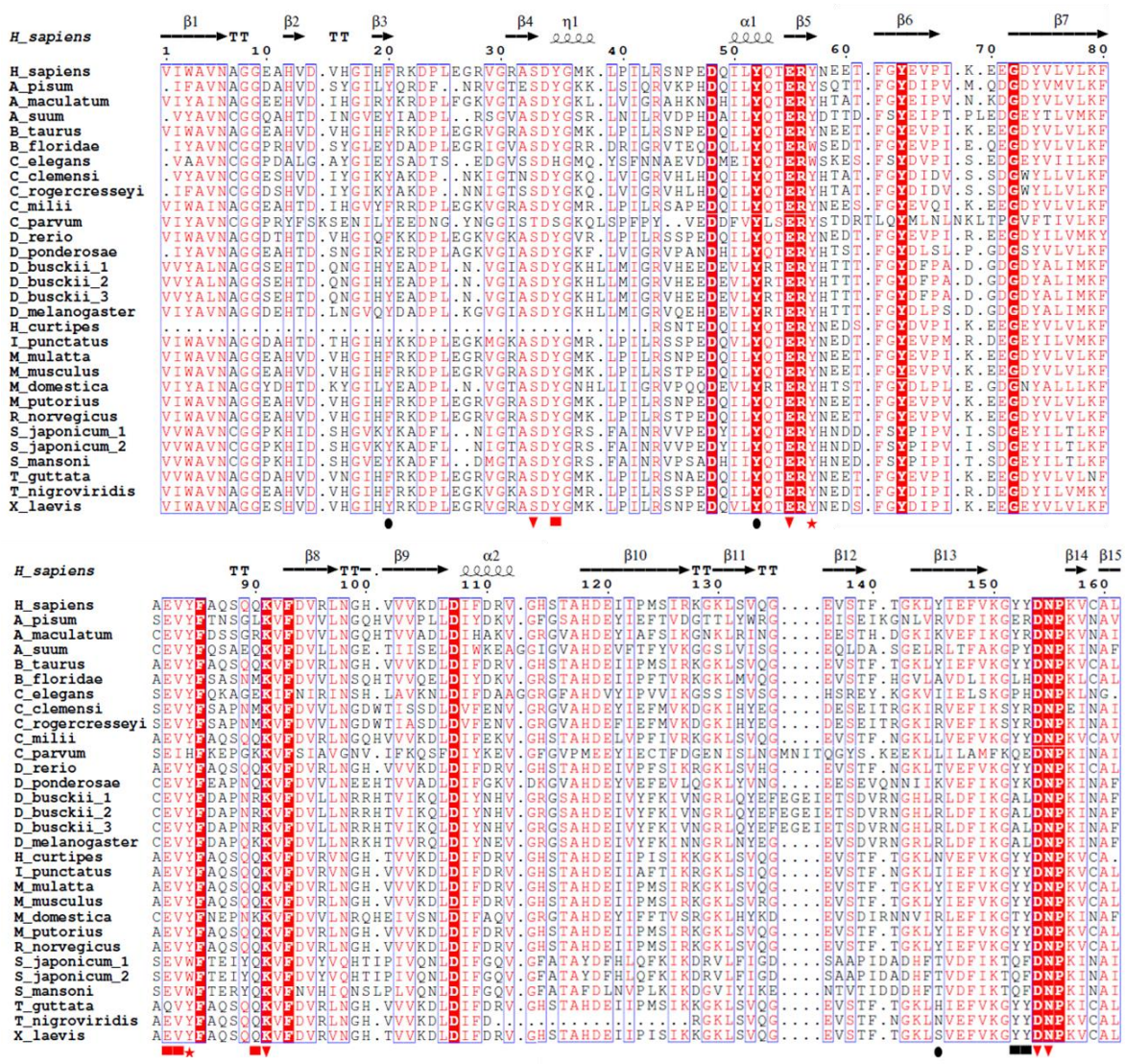


Figure 2.5- Alignment of each malectin-like amino acid sequence in eukaryotes (excluding plants) compared to the human malectin, using ClustalOmega [Sievers, *et al*, 2011] and ESPript [Robert, *et al*, 2014] programs. The amino acids in red are conserved. The red symbols above the alignment represents the binding sites residues; Red triangles-by direct hydrogen bonds; Red squares-by hydrogen bonds mediated by water; red stars- by π/CH interactions. The black symbols mark the carbohydrate-interacting residues from malectin putative binding site.

The conservation among sequences of malectin-like modules is very high. The tertiary structure is conserved, except for 4 modules that have an extension of amino acids between β-sheets 11 and 12. The conservation of the amino acids involved in carbohydrate recognition is extremely high. Exceptions are for *Caeorhabditis elegans* and *Cryptospodidium parvum* species.

C. elegans from Animalia Kingdom has the Tyr35 replaced by histidine. This residue is polar with positive charge and can as well to maintain the hydrogen bond with the carbohydrate. In addition, it is argued that Histidine also has an aromatic chain [Hudson, *et al*, 2015], suggesting that the specificity in this module is the same as the human malectin.

C.parvum from SAR Kingdom has 2 substituted residues: Tyr35 and Tyr84 replaced by Ser35 and His84. Serine has an uncharged polar chain and, although weaker, capable of making hydrogen bonds with the carbohydrate. However, the substitution of Tyr84 by histidine may reflect the loss of π-CH interactions, but a formation of a new hydrogen bond. This alignment result may indicate the malectin-like for this specie has a different carbohydrate specificity, but due to the high conservation in the other interacting residues, it may still participate in the N-glycosylation pathway.

Some aromatic amino acids in the human malectin x-ray structure were observed to be exposed to the solvent and identified as putative binding sites 1 and 2. In the amino acid alignment, the putative binding site 1 includes 3 Phenylalanines at positions 20, 52 and 145. Except for the Phe145, the putative binding site is conserved and thus may be involved in carbohydrate interactions. On the other hand, the putative binding-site 2 with Phe152 and Phe153 shows a lower conservation level than the other putative binding-site.

To better understand the specification of malectin (excluding plants), a phylogenetic tree was constructed (figure 2.6), using the previous alignment.

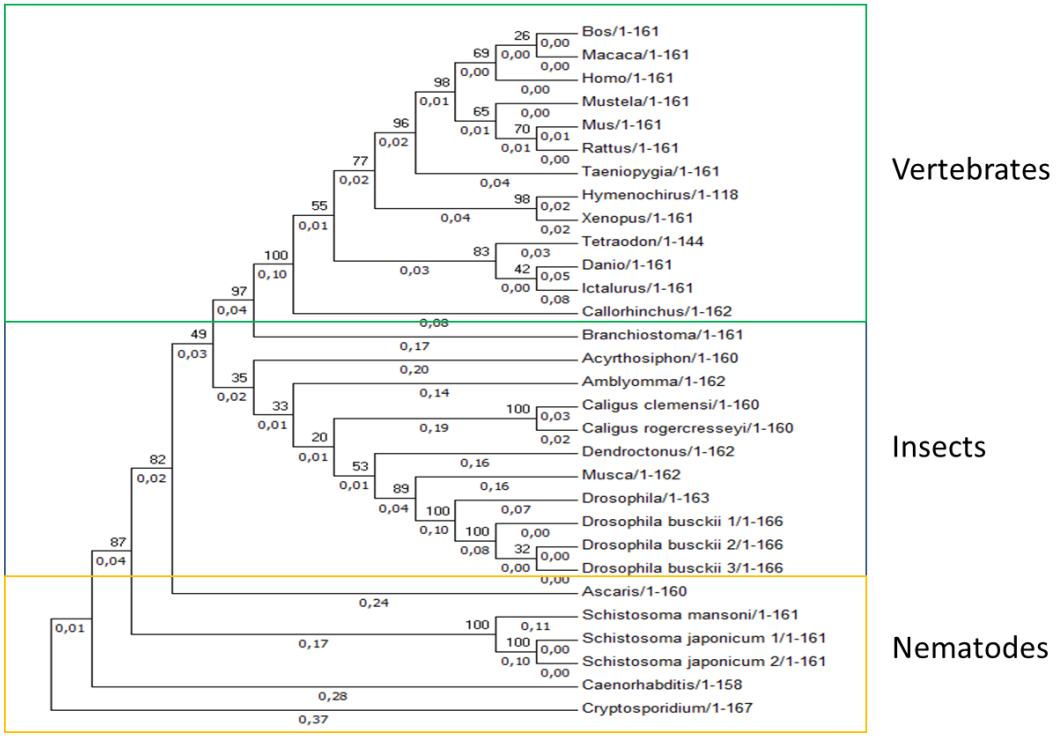


Figure 2.6- Phylogenetic tree construction for all malectin-like in eukaryotes (excluding plants) using MEGA7 program. The method used was Neighbor-joining. Gaps presented in amino acids sequences were treated using the pairwise-deletion option. The validation of phylogenetic tree was done by performing 1000 bootstrap replications.

The *Caenorhabditis elegans* and *Cryptosporidium parvum* are the most distant taxa, which according to the alignments is due to the not conservation of the interacting residues. Moreover, the separation between *Caenorhabditis elegans* and *Schistosoma japonicum* point to the possibility that these are ancestors and existed before the specification process that occurred for the remaining malectin modules that, like the human-malectin, are expected to recognize di-glucosylated oligosaccharides and participate in the N-glycosylation pathways.

2.3.2.2 Plants malectin-like modules evolution:

The sequences of malectin-like modules in plants show a high level of divergence compared to the other malectin eukaryotic modules. For this reason, their alignment is discussed apart from other eukaryotic malectin module (index figure 1).

In comparison to the human malectin sequence, it has an extended N-terminal of a minimum of 20 to a maximum of 200 amino acids residues. However, few exceptions were found. Alignment of the N-terminal in malectin-like modules shows three conserved aromatic residues that may be part of a different interacting site. In addition to the fact that these malectin-like modules are associated with kinases, this suggests that the glycan recognized may be different from the recognized by the human malectin.

Despite the high variance, it was still possible to align their C-terminal part to the human malectin sequence. There are conserved structural elements with human malectin such as α-helice 1 and β-sheets 5, 6, 7, 8 and 14. The interaction residues, on the other hand, aren´t conserved with the human malectin, except for the Glu82 residue. Although the Lys91 in the human malectin has been changed to Arg91 in almost all malectin-like modules, it is a similar amino acid, so carbohydrate interaction maybe maintained.

However, there is some conservation of residues between the malectin-like module in plants. The Glu55, Tyr57 and Asn155 in human malectin are replaced by alanine, valine and glycine, respectably. Both have non-polar chains, thus the carbohydrate interaction in these residues are lost. On the other hand, Gln90 is replaced by a conserved lysine, with a larger and charged chain, capable of making the same type of interaction. This observation together with an extended N-terminal sequence may suggest that the binding-pocket changed of position in the tertiary structure to interact with other type of carbohydrates.

The conservation of the putative biding-sites is identical to the other eukaryotes, except the Phe20 in human malectin doesn´t align with several of these malectin-like modules.

2.3.2.3 Bacteria malectin-like modules evolution:

Malectin-like modules in this group are appended to a catalytic module. A phylogenetic tree (index figure 6) was made to visualize if these malectin-like modules evolved together with the catalytic module from one ancestors only, forming a cluster. Any clustering of related catalytic modules is summarized in table 2.3.

Table 2.3- List of clusters of malectin-like modules associated with a catalytic module.

Domain	Clusters?
Pectin lyase fold	No
Quino protein dehydrogenase	No
Glycoside hydrolase (family 2)	Yes, two clusters
Peptidade S8/S53	Yes, one cluster
TolB-like	Yes, three clusters
PapD-like	No
Galactose oxidase/ Kelch beta propeller	No
Polycystic kidney disease (PKD)	Yes, similar to TolB-like

Malectin-like modules associated with Glycoside Hydrolase family 2 (GH2), peptidase S8/S53, TolB-like and polycystic kidney disease (PKD) have clusters in the phylogenetic tree. The existence of clusters may indicate that the malectin-like modules have co-evolved with the catalytic module, gaining specific carbohydrate specificities. For this reason, 4 alignments of malectin-like modules with their respective associated catalytic module (GH2, Peptidase S8-S53, TolB-like, PKD) have been performed and are discussed below.

Alignment of family 2 of glycoside hydrolases associated Malectin-like modules:

For each cluster, an alignment was performed based on the associated catalytic module. Due to many aligned amino acids sequences, these are shown in index figure 2 to index figure 5.

For this cluster, malectin-like modules in comparison to the human malectin sequence (index figure 2) lost structural elements like β-sheets 12 and 13. On the other hand, there are several modules that have an extension of amino acids between β-sheets: 3-4 and 7-8, although without a high level of conservation. Despite these differences, the other structural elements show a moderate level of conservation, maintaining the β-sandwich structure.

The binding-pocket compared to the human malectin, just Leu82 and Lys91 are conserved, although the lysine is substituted by arginine (a similar residue). Significant changes are observed for Tyr35, Glu55 and Asp154 present in human malectin: several modules have an aromatic amino acid at position 55, whereas the other 2 residues are now glycine. Furthermore, 2 amino acids in β-sheet 4 that aren´t present in human malectin, serine and tryptophan are almost conserved in each malectin-like module and may have carbohydrate interactions. This observation points toward different specificity from the human malectin. In addition, these modules are expected not to share the same carbohydrate specificity.

These modules have 2 of 3 amino acids conserved in a putative binding-site 1. This observation was also verified for eukaryotes. Perhaps, the malectin prior to co-evolution has had the carbohydrate interactions at this site, but there was an adoption of other region (binding-pocket) to increase the plasticity to accommodate other glycans, but maintaining the tertiary structure.

Alignment of Peptidase S8/S53 associated Malectin-like modules:

In this cluster, in comparison to the human malectin sequence (index figure 3), the malectin-like have lost structural elements as β-sheets 11 and 12. Despite that, the structural elements show a high level of conservation.

On the other hand, the alignment of the modules reveals little conservation of amino acids that are predicted to be involved in carbohydrate interactions. Amino acids in human malectin that are responsible for the hydrogen bonds are different, except for sites 82, 83 and 91. In addition, the Tyrosine 57 and 84 in human malectin for π-CH interaction are replaced to other non-conserved amino acids. However, in these modules the sites 36 and 64 are now conserved aromatic amino acids. In this regard, residues for the π-CH interactions have changed, suggesting a different glycan type specificity from the human malectin. It is also suspected each malectin-like module may have different specificities, due to the not conservation of the predicted amino acids for making hydrogen bonds with the carbohydrates.

The conservation of the putative binding-sites is identical as observed to the modules appended to glycoside hydrolase from family 2, except at the position 145 where there is a conserved polar amino acid that can be directly involved in carbohydrate interactions.

Alignment of TolB-like associated Malectin-like modules:

These modules show conservation of structural elements (index figure 4), except for β-sheets 11 and 12 that are absent, while there are some modules that have an extended β-sheet 4.

Malectin-like modules appended to this module showed a higher conservation level of residues predicted (by comparing with the malectin sequence) to be involved in carbohydrate interactions. The principal differences are for Ser35 and Tyr37 (between β-sheets 4 and 5), since the majority of the sequences don't align here with human malectin sequence since there is a gap. However, a high conservation is observed for an aromatic amino acid in the extended β-sheet 4 that may be involved in π-CH interaction.

The TolB-like proteins usually have more than one malectin-module in their sequences. It is anticipated that the modules of the same enzyme may have homogeneous clustering due to the high similarity of predicted interaction residues. However, Tyr57 for π-CH interaction and Gln90 for hydrogen bond in the human malectin aren´t always conserved. Few modules have differences in these residues which could be crucial in slightly altering its specificity. In addition, malectin-like modules appended to this TolB-like domain have high homology and seem to have identical specificities. However, structural characterization is needed to determine the exact specificity.

The conservation of the putative biding-sites, on the other hand, is the same as the module associated with glycoside hydrolase from family 2.

Malectin-like modules alignment associated with PKD alignment:

Like it happened for other clusters, the β-sheets 11 and 12 are missing in comparison to the human malectin sequence and some of the putative interactive residues are conserved (index figure 5).

Several malectin-like modules appended to PKD modules are also associated with TolB-like, so the conservation level is almost similar to observed above, even for the modules that aren´t associated with TolB-like modules.

On the overall, none of the malectin-like modules clustered in the phylogenetic tree that are associated with glycoside hydrolase from family 2 and peptidase S8/S53 and are not expected to have the same specificity as the human malectin since most of the human malectin carbohydrate-interacting residues are not conserved (Ser33, Tyr35, Gln89 and Asp154 - hydrogens bonds and Tyr57 - π-CH bond). Glycoside hydrolases from family 2 have several activities, including of β-galactosidases, β-glucuronidases, β-mannosidases, exo-β-glucosaminidases and endo-β-mannosidase. Hence, it isn´t surprising that the malectin-like modules appended to this catalytic module may have several specificities.

On the other hand, each of the clusters with malectin-like modules appended to either TolB-like or PKD, may have a defined and specific specificities (having perhaps co-evolved with the respective appended catalytic module).

As stated, malectin-like modules appended to glycoside hydrolases from family 2 have shown a higher level of divergence. Due to the currently impact of the human gut microbiome, we will further investigate 4 proteins from the Bacteroidetes phylum.

2.3.3 Analysis of malectin-like modules from two organisms presents in microbiome human: *Bacteroides Ovatus* and *Bacteroides Thetaiotaomicron*:

These species have several unique PULs and homologous module with the malectin sequence (CBM from families 6, 35 and 57) are found in these PULs. The fact that these CBM57 modules come from organisms from different niches in the gut and that they are from different PULs already evidences that they will have different specificities. However, an alignment was carried out to determine the conservation of the putative interacting-binding residues, by comparing to the human malectin sequence.

The selected proteins (BACOVA03100, BACOVA03617 and BACOVA04929 from *Bacteroides Ovatus* and BT0996 from *Bacteroides thetaiotaomicron*) domains were predicted using the InterproScan program [Jones, *et al*, 2014] to identify the CBM57 modules and the associated catalytic modules. The predicted proteins architectures are shown in figure 2.7.

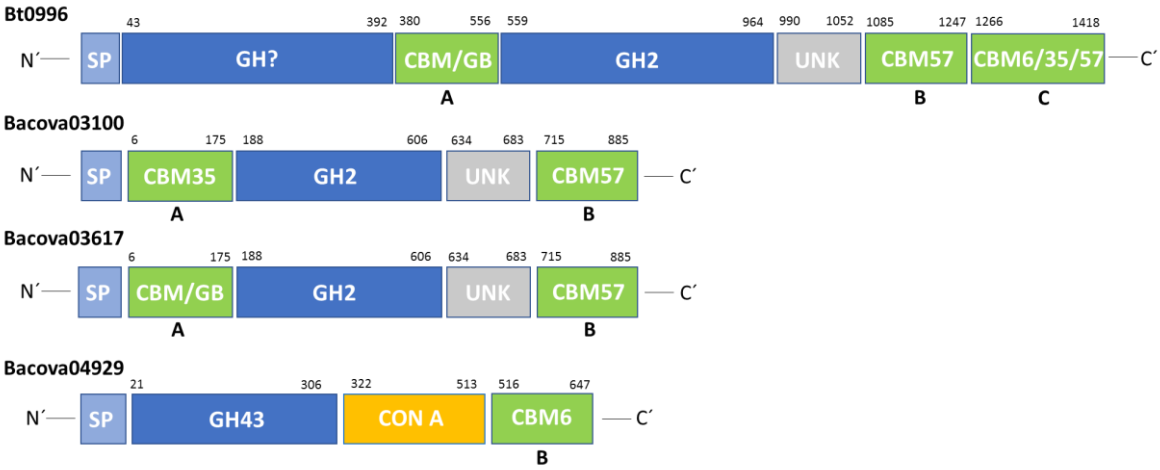


Figure 2.7- Prediction of proteins domains using the InterProScan program [Jones, *et al*, 2014]. SP represents signal peptide. GH2 represents glycoside hydrolase family 2. GH43 represents glycoside hydrolase family 43; CON A represents concanavalin A; UNK represents an unknown function domain. The letters A, B and C represents the modules for the alignment.

The Bt0996, Bacova03100 and Bacova03617 proteins are predicted to have a glycoside hydrolase family 2 (GH2), an unknown domain function, a signal peptide and to contain more than one CBM module. The Bacova03100 and Bacova03617 have a CBM35/GB module, which tertiary structure has high homology to the malectin-like/CBM57 module; while Bt0996 has one CBM/galactose-binding module, a CBM57 module and a CBM module which the program couldn't make the family assignment. This could indicate a possible new CBM family. The Bacova4929 is predicted to have a signal peptide, a glycoside hydrolase family 43 (GH), a concanavalin A and a CBM6 module.

Understanding the glycans specificity of these CBM modules will help to elucidate the activity of the enzyme to which it is appended to. In these three proteins (Bt0996, Bacova03100 and Bacova04929), two or three CBM modules were found. Due to the malectin-like modules of GH2 has shown high divergence level, it is interesting to know if they recognize different glycans, or the same glycan at different branches to direct it to the catalytic module.

So, for the experimental work, we have chosen the CBMs modules from these four proteins for production and purification, to determine the glycan-specificity using glycan microarrays and to characterize them structurally, using x-ray crystallography. These studies have the main purpose of investigating the different glycan-specificities of these malectin-like/ CBM modules and understand their divergent evolution. A second objective is also to understand the degradation mechanism of complex polysaccharides by the PUL system.

2.4 Conclusion:

Bioinformatic analysis is a powerful tool that allows to predict the structure of an unknown protein using only its DNA sequence. This done by looking for homologue proteins already characterized. This allows us to have some insight on its function and biological role.

Malectin-like modules are present in most of the kingdoms of life, with the exception of fungi. In most eukaryotes, the malectin-like is predicted to be individualized and to have a function in N-glycosylation pathway. Two invertebrate species, *C.parvum* and *C.elegans* may have slightly a different glycan-specificity from the human malectin.

In plants and prokaryotes, the malectin-like modules are associated with various catalytic modules. Here, we predict that the modules may have different glycan-specificities from the human malectin. Modules associated with a catalytic module from the same family seem to have different glycan-specificities. However, the modules appended to peptidases S8-S53, TolB-like or PKD modules seem to have a similar and specific glycan-specificity, due to the average high level of amino acids conservation inside the cluster generated for each of these modules.

Since bioinformatics results are mainly predictions, some malectin-like modules from bacteria that seem to have different glycan-specificities from the human malectin were produced (chapter 3) and characterized (chapters 4 and 5).

Chapter 3- Re-cloning, expression tests, production and purification of CBM modules

3.1 Introduction:

In the past, to extract the necessary proteins for structural/catalytic studies or for medical applications, for example, exogenous insulin used for the treatment of diabetes, animals and plants were used to extract the desired proteins. However, that procedure required several large-scale purifications of biological fluids and the amount of final product would still be relatively low [Rosano, et al, 2014].

The technological implementation of the expression of recombinant proteins in prokaryotes (expression proteins that naturally do not exist in the selected prokaryote host) solved the referred problem. In this process, the necessary protein can be expressed in a large quantity and prepared for an easy (usually in one step) purification [Rosano, et al, 2014].

The *Escherichia coli* is the best choice for host microorganism, showing several advantages: 1) a rich media that can be made from accessible components; 2) exogenous DNA transformation is an easy step; 3) the culture has a fast growth [Demain, *et al*, 2009]. These bacterium cells are transformed with a vector that encodes to the recombinant protein. A vector can be a plasmid, a Lambda phage or a chromosomic fragment that is engineering for the production of a heterologous protein [Lodish, *et al*, 2000]. In addition, to select the transformed cells, the plasmid contains a selection marker, which confers resistance to an antibiotic [Lodish, *et al*, 2000]. The promotor used in the expression is usually *lac* promotor, which in presence of lactose induces expression of the heterologous protein. Hence, a large amount of protein is produced [Stevens, *et al*, 2000]. To facilitate the purification process, the recombinant DNA usually codes for an N- or C- terminal tag. The tag adheres to a specific resin column, separating the heterologous protein from the other proteins usually expressed in the host. After purification, if desired, the tag can be removed by chemical or enzymatic cleavage [Stevens, *et al*, 2000]. Tags can have other functions, such as increasing the solubility of the heterologous protein [Rosano, et al, 2014].

Here, we wanted to characterize 7 CBM modules to understand: 1) the divergent evolution of these type-C CBMs; 2) the mechanisms of degradation of complex polysaccharides by the PUL systems. Hence, we needed to first produce the CBM modules, with the process being shown and discussed along this chapter. The adopted strategy consisted of: 1) the re-cloning the sequences of 7 CBMs modules with a C-terminal His-tag; 2) the determination of the best conditions for large-scale recombinant expression; 3) the production and purification of the CBMs modules for characterization studies (chapters 4 and 5).

3.2 Materials and Methods:

The DNA plasmids are part of the Phd work of student Viviana Correia in collaboration of NZYTech company. These DNA plasmids encode CBM57 modules (or homologous- members from CBM families 6 and 35) isolated from two species of Bacteroidetes, *B. ovatus* and *B. thetaiotaomicron*. The information about the recombinant protein sequence, family, theoretical isoelectric point, molecular weight and DNA length are described in table 3.1.

Table 3.1- List of each recombinant protein information, including Locus tag, prevenient organism, CBM family, theoretical isoelectric point, molecular weight and DNA length.

Protein ID/Locus tag	Organism	CBM family	Theoretical pI	Protein Mw (KDa)	DNA Length (bp)
BACOVA03100_A	Bacteroides ovatus	CBM35	5.62	19.38	465
BACOVA03100_B	Bacteroides ovatus	CBM (35, 57)	5.65	19.23	465
BACOVA03617_B	Bacteroides ovatus	CBM57	7.07	21.86	528
BACOVA04929_B	Bacteroides ovatus	CBM6	9.59	16.69	378
BT0996_A	Bacteroides thetaiotaomicron	CBM/GB	6.86	21.98	525
BT0996_B	Bacteroides thetaiotaomicron	CBM57	6.28	20.72	492
BT0996_C	Bacteroides thetaiotaomicron	CBM (6, 35, 57)	9.43	15.89	366

Each recombinant DNA used codes for an additional N-terminal His-tag (with 6 histidines) and has kanamycin resistance.

The His-tag position may influence the protein expression, solubility, purification process and the characterization studies. For the determination of the binding specificity, the His-tag position may block or compete with the ligand, leading to misinterpretation of the data in the analysis. In the structural characterization, the His-tag can be too flexible, preventing the crystallisation of the protein.

The strategy here was the re-cloning of the selected recombinant DNA (HTP constructs) to the pET28 vector with a C-terminal His-tag (with 6 histidines). The first step was to design the primers as well to produce extra recombinant DNA, by transforming *E.coli* cells.

3.2.1 Re-cloning of recombinant DNA into a new vector

3.2.1.1 Recombinant DNA production and isolation

To produce extra recombinant DNA for re-cloning, the competent cells (treated with special protocols for adhering calcium or rubidium to their outer membrane layer) are transformed with the DNA, which the transformation in our lab is performed through heat-chock method. The transformed competent cells are then uniformly spread on medium agar plate with the selected antibiotic and incubated at the desired temperature. Then, an inoculum is made by selecting an isolated colony on the agar plate. For the isolation of the plasmid DNA, the cell culture of the inoculum is centrifugated, followed by alkaline lysis. Afterwards, the plasmid DNA is purified using gel-based silica NZYTech plasmid spin column.

3.2.1.1.1 Transformation:

The following procedure was performed in a sterile environment. At first, 2 µl of the recombinant DNA was added to 50 µl of competent cells. The samples were kept on ice for 30 minutes, followed by heat-chock at 42 °C for 42 seconds and cooled on ice for 2 minutes. In each sample 1 ml of LB medium was added and incubated at 37 °C, 250 rpm for 1 hour and 30 minutes (Orbital ShakerIncubator ES-20, from Grant.bio).

The cells were then collected by a centrifugation for 1 minute at 5000 rpm, being then 950 µl of supernatant discarded. The pellets were re-suspended in the remaining supernatant and spread on the LB agar plate with kanamycin at 50 µg/ml and incubated overnight at 37 °C.

3.2.1.1.2 Inoculation:

In sterile falcons, 10 ml of LB medium and 10 µl of kanamycin at 50 mg/ml were added. A single colony of each transformed recombinant DNA was picked and inoculated into the falcon, followed by the overnight incubation at 37 °C, with 150 rpm shaking (Orbital Shaker Incubator ES-20, from Grant.bio).

3.2.1.1.3 DNA Isolation:

The inoculums were centrifuged at 4000 rpm for 10 minutes to collect the cells (the flow-through was discarded).

DNA extraction and purification was done using the NZYTech miniprep kit, with the following centrifugations done at maximum speed of 13000 rpm.

3.2.1.2 Re-cloning:

Re-cloning the recombinant DNA to other vector requires performing 3 crucial steps: 1) polymerase chain reaction; 2) digestion of the DNA; 3) ligation. The figure 3.1 schematic illustrates the strategy adopted to re-clone the recombinant DNA.

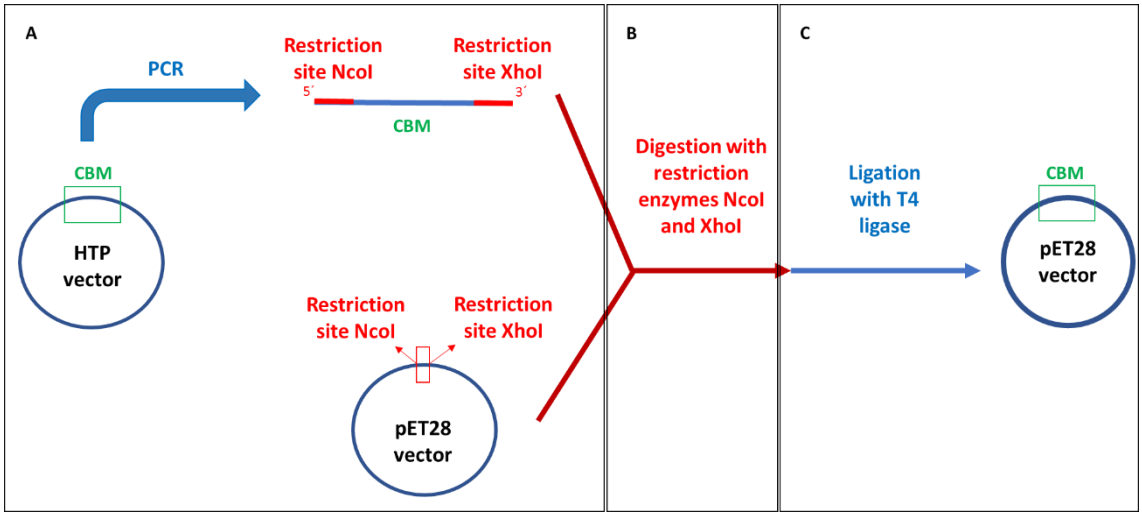


Figure 3.1- Schematic illustration of the re-cloning process. A- Amplification of the DNA recombinant; B- Digestion of DNA with restriction enzymes; C- Ligation of the fragment and plasmid with T4 ligase.

The polymerase chain reaction (PCR) is an enzymatic reaction used for amplifying the DNA fragment of interest, i.e increasing their number of copies and to have the desired restriction sites. To function properly, the PCR requires: 1) primers (short-length fragment that hybridize with complementary DNA sequence); 2) template DNA; 3) mix of nucleotides; 4) a thermal resistance DNA polymerase [Garibyan, *et al*, 2013].

The restriction enzymes are bacterial enzymes that recognize specific sites, known as restriction sites, between 4-8 base pair sequence, which cleave the DNA. For the cloning process, both fragments (PCR product) and vector are digested with restriction enzymes with the aim of both having the same cohesive ends for the ligation reaction.

The enzyme T4 ligase is usually used for the ligation, which has an optical temperature of 20°C. Both the fragment and vector cohesive ends form non-covalent interactions. The T4 ligase covalently joins the cohesive ends by forming of a 3´→5´phosphodiester bond, that requires ATP during the process [Kuhn, *et al*, 2005].

After the incubation time, the competent cells are transformed with the ligated DNA.

3.2.1.2.1 Primers design:

We used the Serial cloner program to design the primers. Various aspects were taken in account. The pET28 vector (index figure 7) has several restriction sites. Here, we choose the DNA fragments to be cloned between NcoI (N-terminal) and XhoI (C-terminal) restriction sites. Thus, the forward primers have a NcoI restriction site and reverse primer a XhoI restriction sites (without stop codon) since we wanted a C-terminal His-tag to be encoded together with the gene encoding the protein of our interest.

The NcoI restriction site is upstream the N-terminal His-tag sequence and the XhoI restriction site is upstream of the region encoding the C-terminal His-tag. When the vector is digested, the sequence encoding the N-terminal His-tag is removed. To the proteins encoding sequences (fragments) are added the restriction sites during amplification with the designed primers. For amplifying the fragments simultaneous, all the primers had a similar melting temperature (Tm) of 57°C. The primers were synthesized at StabVida.

3.2.1.2.2 Amplification of protein encoding fragments:

In each PCR tube, the respective primers and template were added, followed by the addition of the other components necessary for the reaction: Reaction buffer 10X concentrated; $MgCl_2$, the cofactor of DNA polymerase; NZYSpeedyProof DNA polymerase; dNTPs mix and ultrapure water, up to a final volume of 50 µl (listed in table 3.2).

Table 3.2- List of concentrations of each component added in the PCR tubes to a final volume of 50 µl.

Component	Concentration
Reaction buffer 10x	1x
MgCl	2 mM
dNTPs mix	0.4 mM
Primers (Forward and Reverse)	0.8-1.2 mM
Template DNA	10-20 ng/ µl
NZYSpeedyProof DNA polymerase	2.5 U/ µl

A spin was done for each PCR tube before placing in the thermal cycler. The protocol for the performed PCR is described in the table 3.3.

Table 3.3- List of the performed PCR steps, describing the temperature, time and cycles.

Cycle step	Temperature	Time	Cycles
Initial Desnaturation	95ºC	120s	1
Desnaturation	95ºC	45s	30
Annealing	52ºC	45s	
Extension	70ºC	15s	
Final Extension	70ºC	10 min	1

After the PCR was done, an 1.2% agarose gel was run to see the if the PCR product were at the correct size. The PCR products were then extracted from the agarose gel and purified using the NZYGelPure kit.

3.2.1.2.3 Digestion of fragments and vectors:

To each tube, the DNA fragment, the 10X NZYSpeedy Buffer and the restriction enzymes were added as described in table 3.4.

Table 3.4- Description of digestion assay using restriction enzymes

Component	Concentration
10x NZYSpeedy Buffer	1x
DNA	11-34 ng/ml
Speedy XhoI	10 U
Speedy NcoI	10 U

Since the 2 restriction enzymes share the same buffer, the digestion of the fragments and vector was done using both restriction enzymes simultaneously. The enzymatic assay was incubated at 37ºC for 1 hour and 15 minutes. Then, a 1.8% agarose gel was run with the digestion reactions. The DNA bands at the correct size were extracted and purified using the NZYGelPure kit.

3.2.1.2.4 DNA Ligation:

In each tube, 10x Reaction buffer, T4 ligase (10 U/ µl), the vector at a concentration of 20 ng and the fragment at 10-molar excess and ultrapure water were added to a final volume of 20 µl. A solution with every reaction component except the fragments was done to make a control of the ligation reaction. The tubes were incubated overnight at 20 ºC. Then, 50 µl of DH5α cells were transformed with 10 µl of the ligation reaction mixture. The rest of the process is described in section 3.2.1.1.2. The validation of the ligation reaction was done using colony PCR to verify the insertion of the fragment in the vector.

3.2.1.2.5 Colony PCR:

The basis of the colony PCR performed was almost the same as that used for fragments amplification. The exception is that the competent cells were used instead of template DNA [Woodman, *et al*, 2008].

For the isolation of the pure colonies, 3 colonies from each culture plate for each ligation reaction were picked and streaked to new agar plates. The colonies on the new plate were then used for the colony PCR.

In 10 µl of MiliQ water, a small amount of a colony was picked, placed in the tube and re-suspended. Then, these were heated on heat block at 95ºC for 5 minutes, releasing the DNA. The lysate was cooled down on ice and centrifuged at 13000 rpm and the supernatant was collected and added to the PCR tube content.

The protocol for the thermal cycler (table 3.3) was the same except for the melting temperature, which was adjusted to 54 ºC.

The analysis of the amplification was done by loading and running the reaction product on a 1.2 % agarose gel.

The positive colonies were then inoculated for DNA extraction and purification, using the NZYminiprep kit (section 3.2.1.1).

The recombinant DNA extracted was sent and sequenced by STAB VIDA company, to ensure that the DNA samples didn´t have any mutation.

Mutations weren´t observed and expressions tests were performed for the two constructs: pHTP and pET28.

3.2.2 Expression tests:

Expression tests consist of optimizing the recombinant protein product, soluble and with the correct folding. These tests are first done in a small-scale, where several conditions are varied. The conditions are listed in table 3.5.

Table 3.5- List of different conditions used in the expression tests of each recombinant protein in study.

E.coli Strain	Temperature	Induction time	IPTG concentration
BL21	19°C	overnight	0,5 mM
	37°C	3 and 5 hours	
Tuner	19°C	overnight	0,5 mM
	37°C	3 and 5 hours	0,2 mM

The BL21 strain is the most used *E. coli* strain in our lab and thus it was the first strain to be used for testing recombinant protein expression conditions for our proteins. This strain contains the mutation of outer membrane protease OmpT, preventing the degradation of the recombinant protein when the lysis is proceeded. In addition, this same cell line has a deletion of the Lon protease gene, responsible for the degradation of several foreign proteins [Rosano, et al, 2014.

Occasionally, the induction of the expression of the recombinant protein is so enhanced that the competent cells enter in stress, resulting in a misfolded recombinant protein, that precipitates forming inclusion bodies. The Tuner strain is another *E. coli* strain option used for the expression of recombinant proteins. It is a derivate of the BL21 strain, with the difference that it has a mutation of permease lacZ, making it inactive [Rosano, et al, 2014]. Thus, the uptake of Isopropyl β-D-1-thiogalactopyranoside (IPTG), a lactose analogue, is slower [Fernández-Castané,*et al*, 2012].

Other conditions can be used to ensure that the recombinant protein is expressed with the correct folding, as it is listed in table 3.5.

Grown the cell culture, to extract the recombinant protein it is important to break the cell walls. The cells are re-suspended with an appropriate buffer (so that after lysis, the protein isn´t denatured) and ultrasounds are applied for disrupting the cells walls.

To determine if the protein was expressed and in the soluble form, each fraction is then prepared for polyacrylamide gel electrophoresis (SDS-PAGE), added loading buffer and heated to boiling temperature (100ºC) [Roy, *et al*, 2014].

For the visualization of the recombinant protein, the gel must be dyed and de-stained for visualizing the recombinant protein. By comparing with the molecular marker sample, it is observed if the protein is present in the estimated size.

3.2.2.1 Protocol for expression with IPTG-induction

In day 1, 50 µl of competent cells (BL21 and Tuner strains) were transformed with 2 µl of recombinant DNA (transformation protocol described in section 3.2.1.1.1).

In day 2, pre-inoculum was made by picking and putting a colony in 10 ml of LB medium with 50 mg/ml of kanamycin and incubating at 37ºC.

In day 3, 100 µl from the pre-inoculum was added to 10 ml of LB medium with 50 mg/ml of kanamycin. Each cell culture was incubated at 37ºC at 180 rpm (Shaker IS-971R from Lab Companion) until OD_{600nm} reached 0.5-0.8, followed by the addition of 0.2 or 0.5 mM of IPTG. The expression induction was then carried out at the desired temperature and induction time, with a shaking of 180 rpm.

3.2.2.2 Cell harvesting and lysis:

The pellets were re-suspended with 1 ml of buffer: 50mM HEPES, 5mM $CaCl_2$, 1M NaCl and 10mM of imidazole. Then, the cells were disrupted using a sonicator (UP100H, from Hielsher) with 1 cycle of 1 minute, and then the lysate was centrifuged at 13000 rpm, 4ºC, for 30 minutes. Both flow-through and pellets were collected (the pellet was re-suspended with 1 ml of buffer).

3.2.2.3 Analysis by polyacrylamide gel electrophoresis (SDS-PAGE):

To perform the SDS-PAGE, 25 µl of each sample were mixed with 5 µl of 6X loading buffer. The samples were then heated on a heat block at 100 °C for 5 minutes, followed by 2 minutes of incubation on ice and finally briefly centrifuged for collecting the entire sample. In the SDS-PAGE, 15 µl of each sample were applied onto the gel wells and the running was performed at 200 V and 200 mA for 1 hour. At the end of the run, each gel was stained with Coomassie blue for 15 minutes, and de-stained for 30 minutes with a mixture of 40 % methanol and 15% of acetic acid in water, being ready for visualisation.

3.2.3 Growth in large scale and purification:

Selected the favourable conditions for the recombinant protein expression of our proteins, the conditions were replicated for a larger-scale protein production of 1 L of cell culture (figure 3.2).

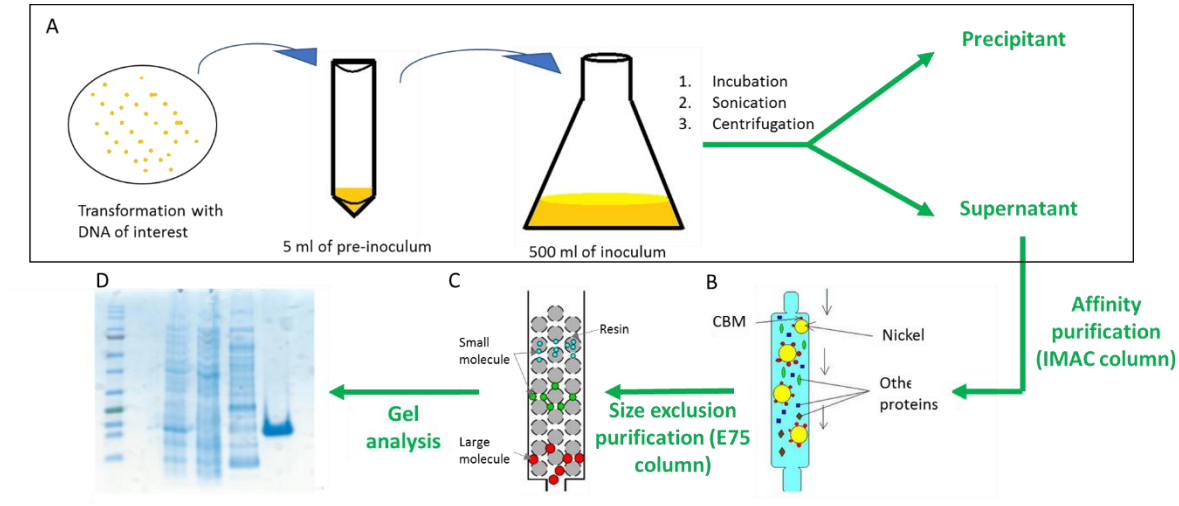


Figure 3.2- Schematic illustration of production and purification of our 2 CBM modules. A-production and extraction of the CBMs. The cell culture growth of 1 L was performed using 2 Erlenmeyer flask containing 500 ml of inoculum. B- Purification of the modules using affinity chromatograph. C- Occasionally, the size exclusion chromatograph was aided to increase the purity of the sample. D-Analyse of the sample by either SDS-PAGE and NATIVE-PAGE, for further characterization studies.

The extraction of the recombinant protein was to follow the protocol in section 3.2.2.2, with few exceptions. In this case, per 1g of pellet, 10 ml of buffer were added to re-suspend the cells before the lysis. Furthermore, the sonication cycles were increased to 5.

After centrifugation, just the supernatant was collected for the purification of CBM modules, by affinity chromatography.

In immobilized-metal affinity chromatography (IMAC), the histidine residues bind strongly to bivalent metal ions such as Co^{2+} or Ni^{2+}, immobilized in the column resin [Bornhorst, *et al*, 2000]. The His-tagged CBM module binds to the column while others bacterial proteins are eluted and removed from our sample. To elute the CBM module, a buffer containing a higher concentration of imidazole is used, which competes with the His-tags for the binding to the metal ion and the module is eluted.

Occasionally, we observed through SDS-PAGE that the recombinant protein needed further purification, using size exclusion chromatography [Mori, *et al*, 2013].

3.2.3.1 Affinity chromatography:

Each CBM module purification was done using a His Trap™ column of 5 ml Ni^{2+}, attached to the chromatograph ÄKTA START (both from GE-Healthcare), with the UNICORN™ start 1.0 control software.

Two different buffers were used for the imidazole gradient: Buffer A- 50 mM HEPES, 5 mM $CaCl_2$, 1M NaCl and 10 mM of imidazole, at pH of 7.5; Buffer B- 50 mM HEPES, 5 mM $CaCl_2$, 1M NaCl and 500 mM of imidazole, at pH of 7.5. Prior to purification, the column was washed with 50 ml of Mili.Q water and equilibrated with 20 ml of buffer A.

The soluble fraction (cell extract) was then loaded into the column. The content that is loaded into the column is monitored by a UV cell at 280 nm.

After the elution of bacterial proteins, a first wash with 10% of buffer B was done to remove the remaining bacterial proteins that could bind with low affinity to the column. In 100 ml of flow-through, a gradient of imidazole was then made from 10 to 100%, to elute the CBM module.

Collected the fractions, they were analysed by SDS-PAGE and Native PAGE for the confirmation of protein purity and stability (the latter if possible, due to the protein isoelectric point).

3.2.3.2 Analysis by native polyacrylamide gel electrophoresis (Native-PAGE):

The Native-PAGE has the purpose of evaluating the protein stability and number of conformations present in solution (rather than the size). Regarding to this, several solutions used (to treat the sample and to make gels) are changed to not denature the protein, in order to maintain the conformation. The amounts used for the visualization of the gel were the same as used in the SDS-PAGE [Hong, *et al*, 2012].

3.2.3.3 Size exclusion chromatography:

Size exclusion chromatography has the principle in which molecules in solution are separated by their molecular weight. A resin with pores of different sizes is used. Molecules with higher molecular weight cannot enter into the pores, thus their elution is faster. On the other hand, molecules with lowest molecular weight can enter into several pores, requiring more time to be eluted.

Here, we used a Superdex 75 column (GE-Healthcare) coupled with the chromatograph SHIMADZU Corporation. The column was washed with 50 ml of MILI-Q water, equilibrated with 50 ml of desalting buffer (section 3.2.3.4). The purification was performed by injecting 1ml of our sample at a time. The run was monitored using a UV cell at 280 nm.

Each fraction was collected and analysed by SDS-PAGE.

3.2.3.4 CBM modules Desalting and Concentration:

It is crucial to desalinate the sample, otherwise the imidazole may influence the crystallisation assay. A technique for desalting is to change the protein buffer (buffer exchange), using a column with resin Sephadex G-25, to separate small molecules (salt) from our CBM modules (an example in the figure 3.3).

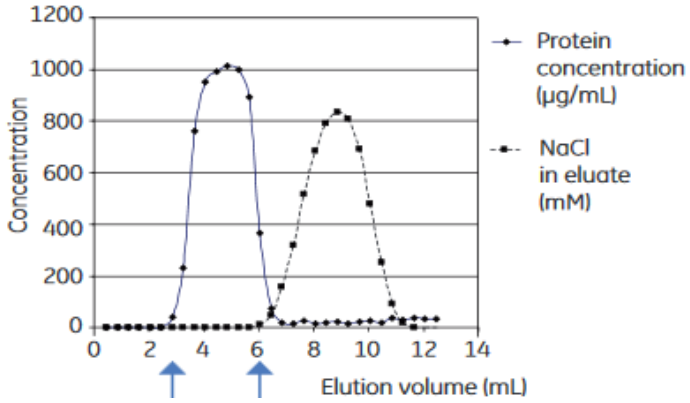


Figure 3.3- An example of desalting chromatogram. There is a clearance of NaCl from the protein when the elution volume is between 2,5-6 ml. To obtain high concentration of protein, 3,5 ml of elution volume should be used. After 6 ml, NaCl begins to be eluted.

Four HiTrap Desalting columns were coupled to the chromatograph ÄKTA START, with maximum of sample volumes of 6 ml at a time. The CBM modules were buffer exchanged to 50 mM HEPES, 5 mM $CaCl_2$ and 100 mM NaCl.

After the elution, the CBM modules were ready for the characterization of their binding-specificity (chapter 4). For the crystallisation assays (chapter 5), the concentration was increased to concentrations of 20-30 mg/ml, using a concentrator with 10 kDa molecular-weight cut-off.

3.3 Results and Discussion:

3.3.1 Re-cloning of CBM modules:

More quantity of the HTP recombinant DNAs was obtained using a miniprep protocol from NZYTech. The sequence encoding the CBMs of interest were amplified by PCR. The PCR products were analysed by agarose gel electrophoresis (figure 3.4). Their lengths are detail in table 3.1.

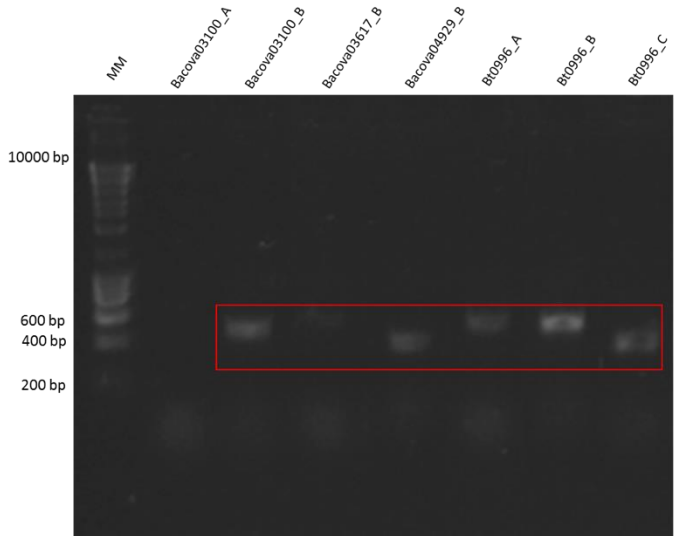


Figure 3.4- Agarose gel (1.8%) electrophoresis intercalated with Safe Red of 7 PCR products from CBM modules of two Bacteroidetes species. The bands of amplified PCR products are highlighted with the red rectangle. MM-NZYDNA Ladder III.

The PCR amplification of the recombinant DNAs was successful for most of the clones, with the exception of the Bacova03100_A DNA. One of the reasons could be the not correct annealing temperature. In addition, the Bacova03617_B DNA was poorly amplified, being also excluded.

The PCR products were cleaned up using the NZYGelpure protocol and digested, like the pET28 vector, with Nco I and Xho I restriction enzymes. The result of the digestions was analysed (figure 3.5) and purified from the agarose gel, using NZYGelpure protocol.

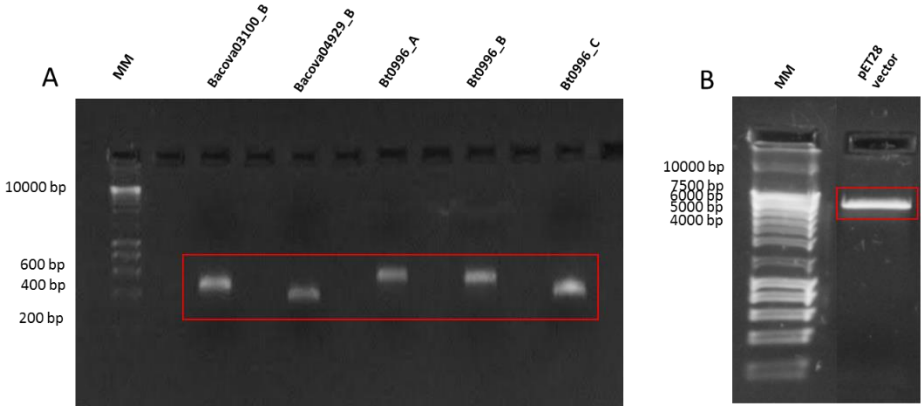


Figure 3.5- Agarose gel (1.8%) electrophoresis intercalated with Safe Red of 5 fragments (A) and the pET28 vector digestion (B). The bands of digested fragments and pET28 vector are highlighted with the red rectangle. MM-NZYDNA Ladder III.

Isolated the DNAs from the gel, the concentrations were measured using plate reader (Molecular Devices) at 260 nm (around 10-30 ng/µl), and the ligation was performed using the T4 DNA ligase protocol, followed by transformation.

To determine the success of the ligation, 2 colonies were tested using colony PCR. To visualize the amplification of the gene of interest, the reaction volumes were analysed in the agarose gel. Initially, we had problems in optimizing the colony PCR. Thus, the analysis of the re-cloning of 2 CBM modules were done by a more traditional method that consisted in the: extraction of the DNAs using miniprep, digestion of the DNAs and analyse of the band sizes on the agarose gel. This method required more steps and time.

The agarose gels for both methods are shown in figure 3.6.

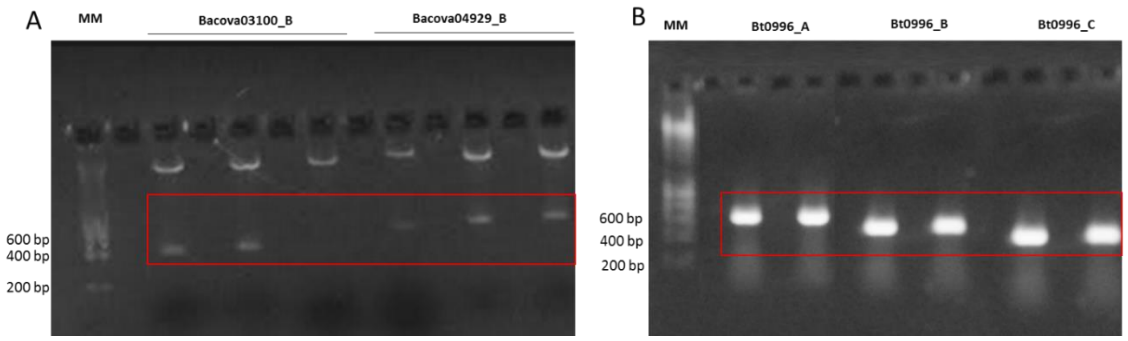


Figure 3.6- Agarose gel (1.8%) electrophoresis intercalated with Safe Red of the digestion recombinant (A) and the amplification of the fragments (B). The bands of digested recombinant DNAs (A) and amplificated fragments (B) are highlighted with the red rectangle. MM-NZYDNA Ladder III.

By the analysis of the agarose gels, we had 5 CBM modules inserted in the pET28 vector. Each of these DNA were sent for sequencing at the STAB VIDA company and the DNAs sequences confirmed.

Having re-cloned 5 CBMs modules from a total of 7, we performed the expression tests, to find the best condition for each of the CBM modules that produced the highest yield in the soluble fraction.

3.3.2 Expression tests:

Here, we subdivided the analysis of the expression tests based on the position of the His-tag: N-terminal His-tag and C-terminal His-tag

3.3.2.1 N-terminal His-tagged CBM modules expression:

As referred in section 3.2.2, several expression tests were performed. Here, we will show the SDS-PAGEs that were observed for 2 CBM modules that expressed and in the soluble fraction (figure 3.7).

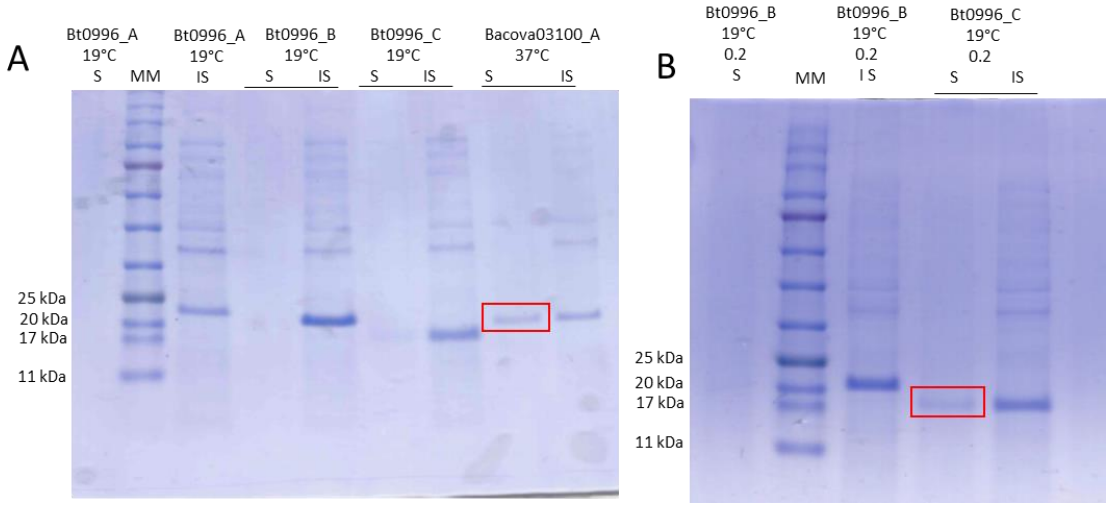


Figure 3.7- SDS-PAGE (10% acrylamide) analysis of the expression of the CBM modules. The bands of the expressed CBM modules in the soluble fraction are highlighted with the red rectangles. MM-NZYDNA Ladder III. S-soluble fraction; IS-insoluble fraction; 0.2 and 0.5-IPTG concentration for induction; Band inside of red rectangle show the proteins that expressed in the soluble form and at the correct size; MM-NZYColour Protein Marker II.

The SDS-PAGE analysis showed that just 2 CBM modules, in a total of 7, expressed in moderate levels in the soluble fraction: the Bacova03100_A and Bt0996_C modules. For Bacova03100_A module, the best condition was using BL21 strain, induced with 0.5 mM of IPTG at 37ºC for 3h. For Bt0996_C modules, the best condition was using Tuner strain, induced with 0.2 mM of IPTG at 37ºC for 3h.

3.3.2.2 C-terminal His-tagged CBM modules expression:

There wasn´t an improvement of the expression, except for the Bt0996_C module, shown in figure 3.8.

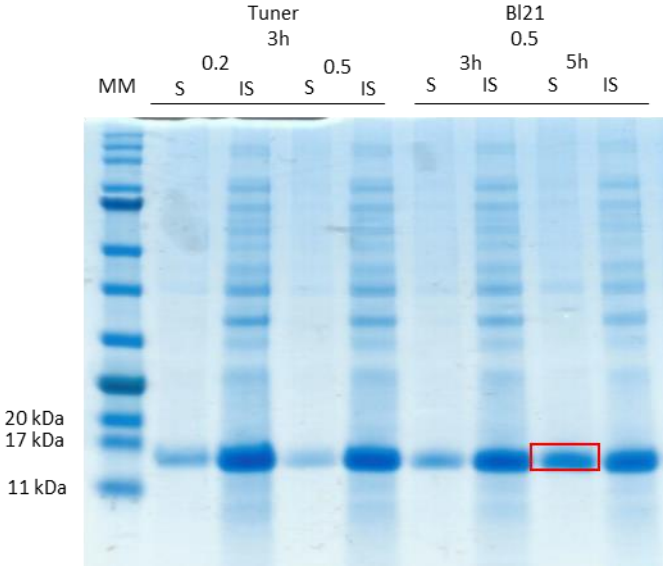


Figure 3.8- SDS-PAGE (10% acrylamide) analysis of the expression of the Bt0996_C module with C-terminal His-tag, induced at 37ºC. The bands of the expressed CBM module in soluble fraction is highlighted with the red rectangles. MM-NZYDNA Ladder III. S-soluble fraction; IS-insoluble fraction; 0.2 and 0.5-IPTG concentration for induction; Band inside of red rectangle is one of the protein that expressed in soluble form at the correct size; MM-NZYColour Protein Marker II.

The changes of the His-tag position, from N-terminal to the C-terminal, increased its level of expression in a different expression condition: Bl21 strain, concentration of IPTG of 0.5 mM, 5 hours of induction at 37ºC.

Since the His-tag position can have influence in the purification process or interfere with the characterization studies, the strategy adopted was to produce in large-scale both constructs of Bt0996_C module.

The best conditions for large-scale expression of the 2 CBM modules are summarized in the table 3.6.

Table 3.6- List of the best conditions for 2 CBM modules expression.

Protein	*E.coli* strain	Temperature	Induction time	IPTG concentration
Bacova03100_A	BL21	37ºC	3h	0.5 mM
His_Bt0996_C	Tuner	37ºC	3h	0.2 mM
Bt0996_C_His	BL21	37ºC	5h	0.5 mM

For the other CBM modules which were expressed in insoluble form, other expression could be tested for future work.

These CBM modules are part of a protein and their current constructs may be unstable. Other tags that confer stability could be used.

3.3.3 Large scale expression and purification:

In this section, we will show all steps involved in purification of the CBM modules.

3.3.3.1 Bacova03100_A module purification:

Since this module has 2 cysteines in its sequence, we will present results from 2 purifications, one without dithiothreitol (DTT), and other with 10 mM of DTT, a reducing agent, added to the purification buffer A (detailed in section 3.2.2.2). Chromatograms of IMAC purifications are shown in figure 3.9.

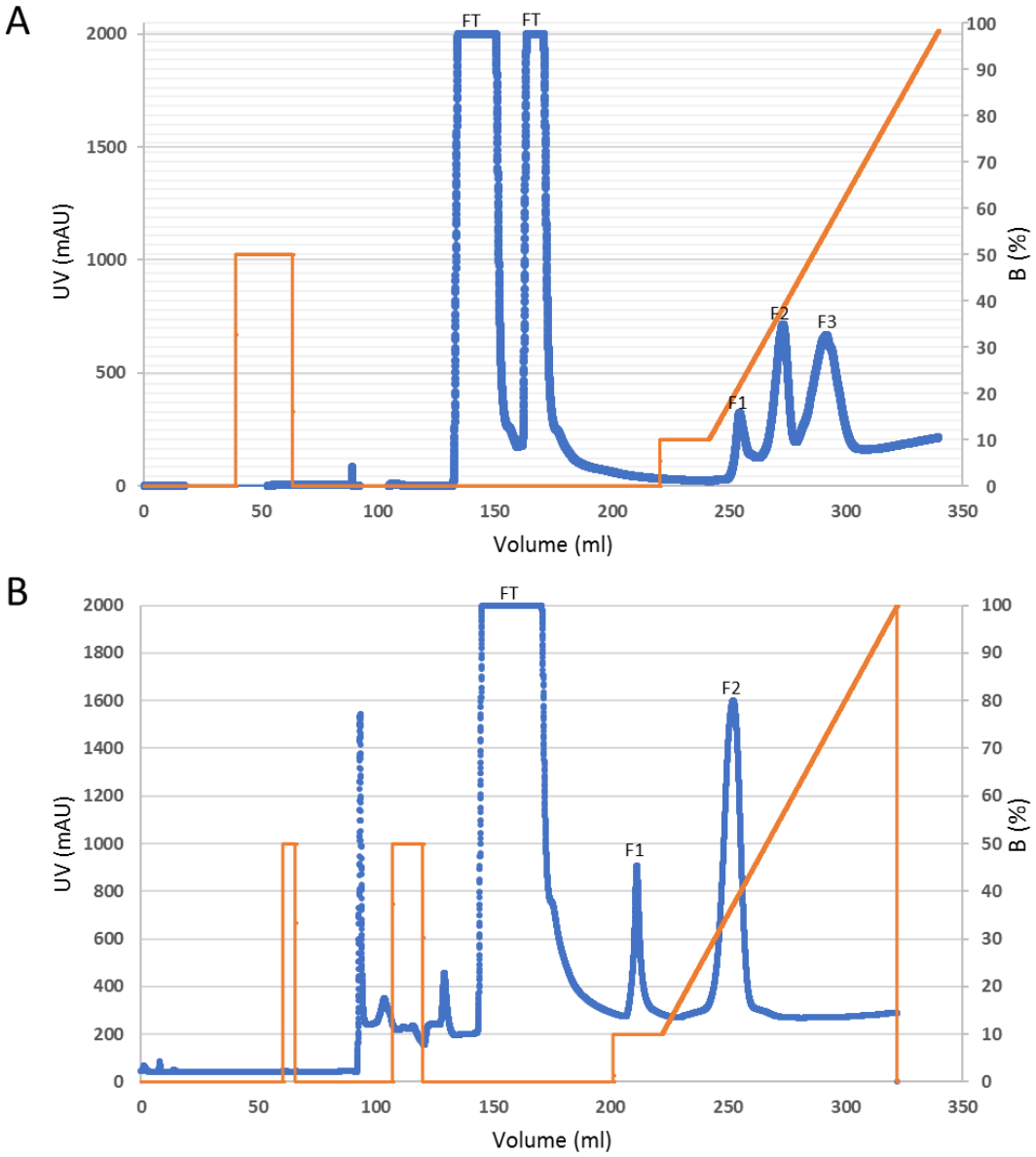


Figure 3.9- Results from the purifications of Bacova03100_A. IMAC Affinity chromatogram without (A) and with DTT (B). Blue line represents the volume as function of UV. Orange line represents the function of a gradient of buffer B; FT-flow through, F1-fraction eluted with: A-10% of buffer B (60 mM imidazole); F2-fraction eluted around 30% of buffer B (160 mM imidazole), F3-fraction eluted around 50% of buffer B (260 mM imidazole).

Without DTT (chromatogram A), fractions from 3 peaks were collected, while with DTT (chromatogram B) fractions from 2 peaks were collected during the imidazole gradient. All collected fractions were analysed by SDS- and Native-PAGE, to ensure that the fractions used in the following studies had the CBM module pure and well-folded (figure 3.10)

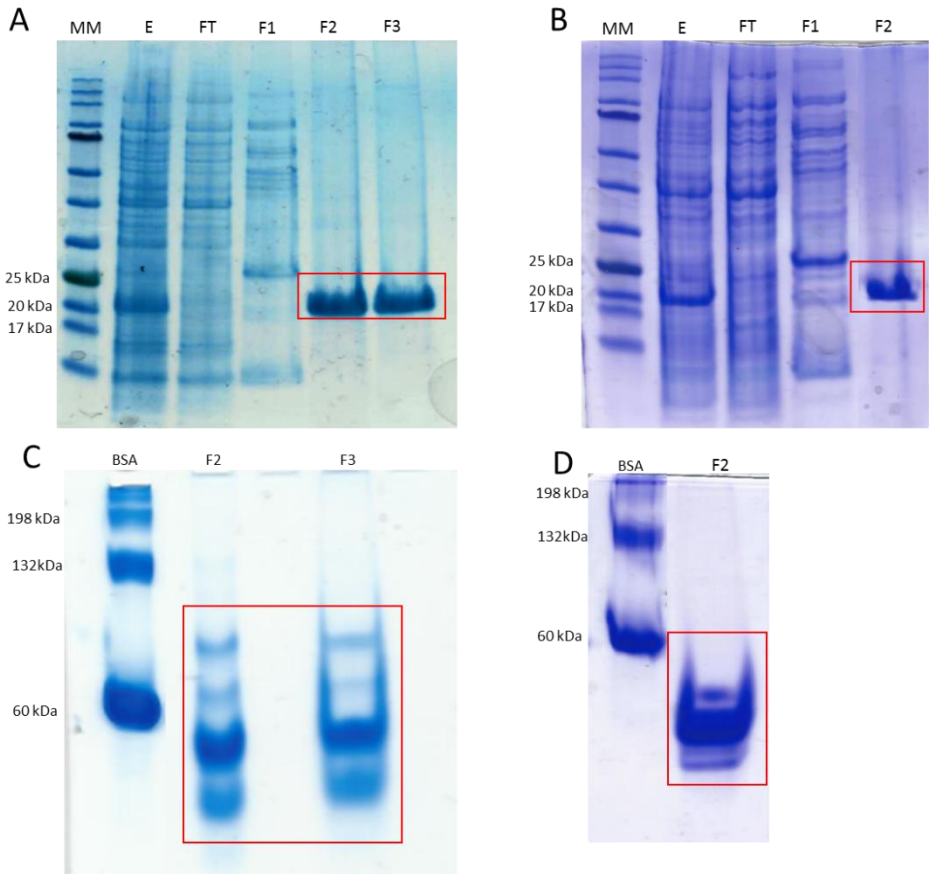


Figure 3.10- SDS-PAGE (10% acrylamide) analysis of IMAC purification without (A) and with DTT (B); Native-PAGE (12.5% acrylamide) of the IMAC purification without (C) and with DTT(D). The bands of Bacova03100_A modules are highlighted with the red rectangle; FT-flow through, F1-fraction eluted with 10% gradient; F2-fraction eluted with 30% gradient; F3-fraction eluted with 50% gradient; MM- NZYColour Protein Marker II.

It seemed, by the SDS-PAGE analysis, that we had obtained pure fractions of the Bacova03100_A module. However, by analysis of Native-PAGE, the sample without the addition of DTT (figure 3.10, image C) showed 4 isoforms in the gel, not being appropriate for characterization studies. On the other hand, the sample with DTT showed almost exclusively a single band (figure 3.10, image D) and so it proceeded for characterization studies.

This fraction was desalted. For structural characterization studies, the protein was concentrated using concentrators Vivaspin 10KDa (GE Healthcare), until reaching a concentration of 20-30 mg/ml.

3.3.3.2 Bt0996_C module purification:

For this module, we had 2 constructs with different positions of the His-tag that were successfully expressed. Large-scale grown was performed for the two constructs and the chromatograms and their respective SDS-PAGE analysis are shown in figure 3.11.

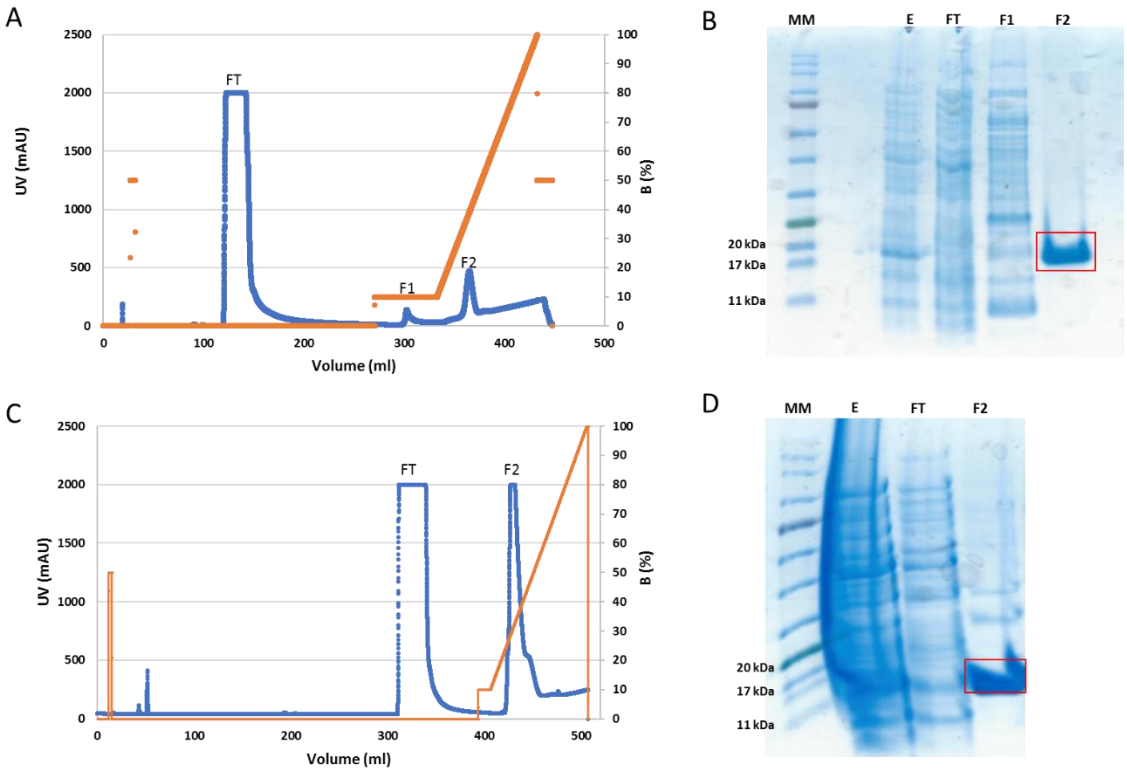


Figure 3.11- A and B- Results from purification of Bt0996_C, with N-terminal His-tag, showing the IMAC affinity chromatogram (A) and the SDS-PAGE (10% acrylamide) (B); C and D- Results from purification of Bt0996_C, with C-terminal His-tag, showing the IMAC affinity chromatogram (C) and the SDS-PAGE (10% acrylamide) (D); Blue line represents the volume as function of UV. Orange line represents the function of a gradient of buffer; The bands of Bt0996_C are highlighted with the red rectangle; E-extract cells before purification; FT-flow through, F1-fraction eluted with 10% of buffer B (60 mM Imidazole); F2-fraction eluted around 30% of buffer B (160 mM Imidazole); MM-NZYColour Protein Marker II.

We obtained with high purity the Bt0996_C module with N-terminal His-tag sample, proceeded to the desalting protocol and concentrated for characterization studies. The Bt0996_C module with C-terminal His-tag needed a second purification step. This was performed using a size exclusion chromatography with an S75 column (GE Healthcare). The respective chromatogram and SDS-PAGE are shown in figure 3.12.

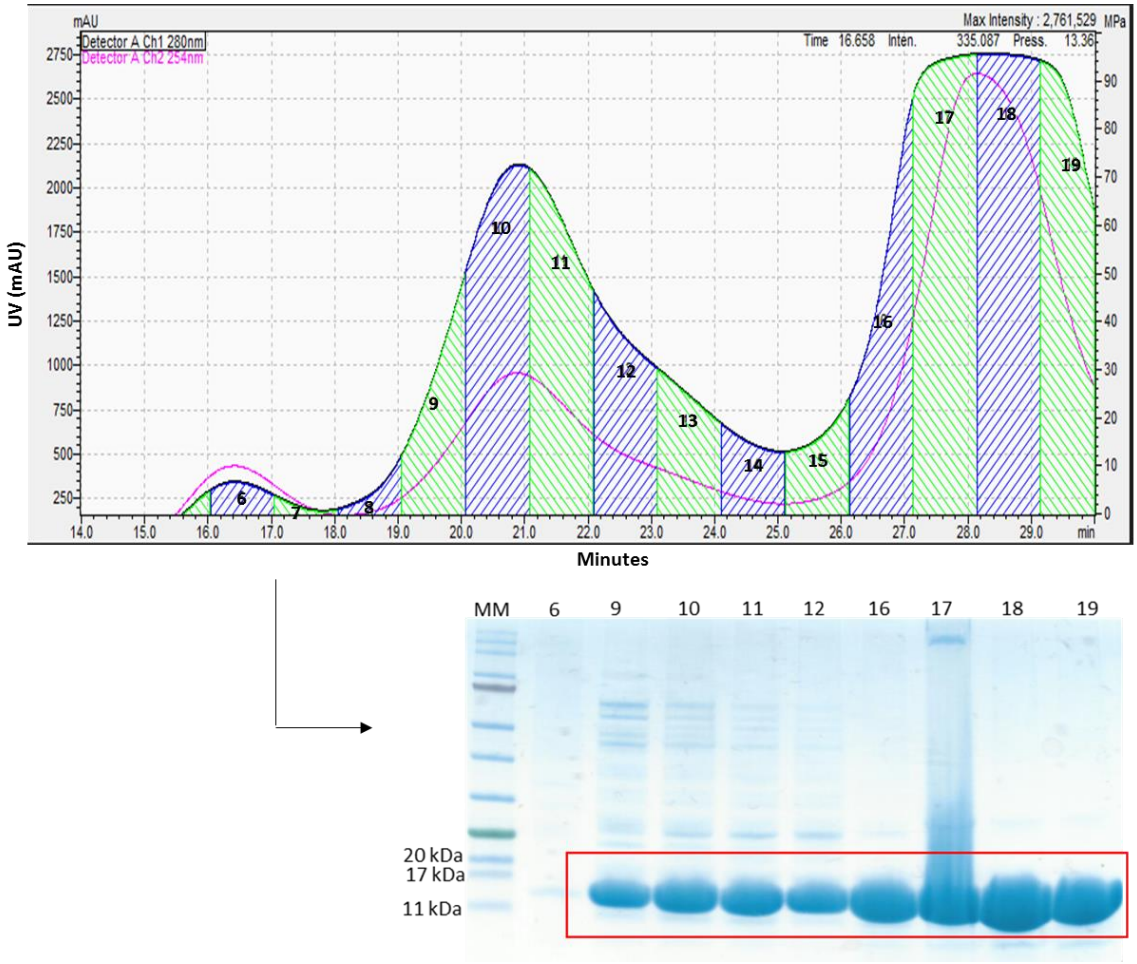


Figure 3.12- Result from purification of Bt0996_C-His, by size exclusion chromatography using E75 column. The chromatogram represents the time (minutes) as function of UV. The sample was fractionated and each promising fraction was analysed with 10% SDS-PAGE.

We obtained pure Bt0996_C-His module within fractions 16, 17, 18 and 19, that were collected and concentrated for characterization studies. It should be noted that fractions 9, 10, 11 and 12 had our module with other bacterial proteins. The existence of the module in these fractions could be due to dimerization, which would thus increase the molecular weight of the molecules and making them to be eluted first, along with other bacterial proteins.

As this module has a theoretical isoelectric point of 9.43, a regular Native-PAGE gel couldn´t be used for the analysis of these fractions. The CBM module was concentrated around of 20-30 mg/ml, to be used in further studies.

3.4 Conclusions:

The constructs and vectors used, the expression and induction conditions have a great impact in the protein expression.

Beginning with 7 CBM modules, we could only successfully obtain 2 modules in the soluble fraction: Bacova03100_A and Bt0996_C. For the Bt0996-C module, two constructs (pHTP and pET28) expressed. Although the pET28 construct had increased levels of expression, the IMAC affinity purification wasn´t so effective.

Chapter 4- Specificity characterization of CBMs using glycan microarrays

4.1 Introduction:

Glycan microarrays have been developing during various years and implemented for the determination of glycan-binding specificities of several proteins [Heimburg-Molinaro, *et al*, 2009]. It is a high-throughput technique that enables the simultaneous determination of the specificity of several proteins to different carbohydrates. This is done by probing the proteins against various structurally diverse glycans, using low amounts of both biomolecules.

To applicate the glycan microarrays technique, there are diverse glycan libraries sets (of related glycans) that can be used. Examples of these glycans libraries are presented in the table 4.1. These libraries have the information of the number of samples, the research group responsible, the glycans nature and the immobilization method [Rillahan, *et al*, 2011].

Table 4.1- Description of glycan libraries examples used in glycan microarrays assays. It is presented the number of glycans that compose the library, the research group, the source of the glycans and the immobilization method.

Glycan Type	Number of samples	Research Group	Synthetic	Naturally Derived	Immobilization method
Mammalian	500	CFG	yes	yes	Non-covalent
	600	Feizi	yes	yes	Covalent
	200	Bovin	yes	no	Non-covalent
	200	Cummings	no	yes	Non-covalent
	200	Gildersleeve	yes	yes	Covalent
Bacterial	96	CFG	no	yes	Non-covalent
	48	Wang	no	yes	Covalent

4.1.1 Glycan sources:

As already stated, glycans used in these libraries can be from different sources: natural or synthetic, by chemical or chemo-enzymatic reactions.

Natural glycans sources were the first to be printed in the glycan microarrays and still feature in diverse glycan libraries [Rillahan, *et al*, 2011]. Glycans from different natural sources include polysaccharides from bacteria and plants, milk oligosaccharides and glycans released from cell walls by endoglycosidases or chemical hydrolysis [Varki, *et al*, 2009]. However, the major challenges of using glycans from natural sources are their structure determination (monosaccharides have very similar structures and polysaccharide have similar mass weighs) and their extraction from pure fractions [Rillahan, *et al*, 2011]. In nature, it is relatively difficult to obtain a homogenous amount of the pure glycan.

Other approach is to use synthetic glycans. The principle is to conjugate two monomers by the anomeric hydroxyl group on the ring, either in alpha or beta linkage. The chemical synthesis approach requires the use of complex blocking compounds to protect the hydroxyl groups that have an identical reactivity for the desired hydroxyl group that is attacked by the glycosyl acceptor (figure 4.1) [Smoot, *et al*, 2009].

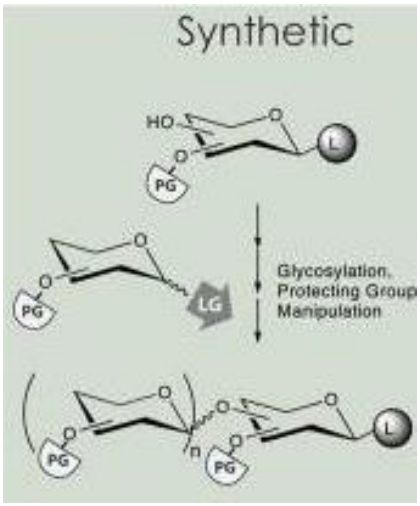


Figure 4.1- Illustration of glycans synthesis. PG represents the protecting group. LG represent the leaving group. Image adopted from article: Glycan microarrays for decoding the glycome.

Although there are approaches where attempts are made for using enzymatic synthesis, there is a limitation of the available glycosyltransferases for the desired specific glycosyl linkages. Although still in development, it has been already combined with the chemical approach [Blixt, *et al*, 2006].

4.1.2 Glycan immobilization types:

In glycan of microarrays assay, it is essential that the glycans are immobilized on the array matrix. This immobilization is performed by robotic instrumentation and is characterized by covalent or non-covalent interactions [Rillahan, *et al*, 2011].

The non-covalent immobilization method is usually done by applying the sample in nitrocellulose [Wang, *et al*, 2002] or oxidized polystyrene membrane [Willats, *et al*, 2002] on the glass slide. The hydrophobic and electrostatic forces that exists in the glycan enable its immobilization in the membrane. However, small glycans don´t have enough non-covalent interactions to be adsorbed in the matrix, and during the assay they would be probably dissolved and removed from the matrix, misleading the data interpretation [Rillahan, *et al*, 2011]. A strategy is to derivatize these glycans with a long-chain alkyl attachment to increase the hydrophobicity glycan nature (being the derivatized glycan now denominated as neoglycolipids) [Liu, *et al*, 2007]. Other strategy of non-covalent immobilization that is in development is to covalently link oligonucleotides into the glycan, giving them the function to adhere to the matrix (containing complimentary DNA) [Rillahan, *et al*, 2011].

In covalent immobilization methods, the glycans are derivatized with thiol or amine-terminated groups. When the glycan is coupled with an amine group, the matrix used is composed of cyanuric chloride that covalently binds to it [Blixt, *et al*, 2004]. Other strategy is the attachment of underivatized glycan onto a matrix containing photoreactive groups [Carroll, *et al*, 2006].

Independent of the immobilization method chosen, the robot instrumentation can be used to print these glycans either by contact or non-contact onto the matrix. In contact printing, the glycans solutions are prepared in a multi-well source plate at the desired concentrations, and a set of steel pins (1-48) dip down in the solutions and are blotted into the matrix glass slide by pressing the set of pins [Heimburg-Molinaro, *et al*, 2009]. The glycans volume printed is about 0.5 nl. For non-contact printing, the robot uses glass capillaries instead of pins to transfer the glycans into the matrix. These are controlled by electric signals. The printing spots are estimated to be at 0.3 nl and although these are more precise (uniformly delivered), the number of tips used are limited to 4 or 8, because of their complexity and expense, being thus required several hours to print slides. Thus, because of the time required, a special attention for the glycans is needed (ex: if the glycans don´t precipitate) [Heimburg-Molinaro, et al, 2009]. For quality control, a fluorophore is usually used for the glycans spots visualization.

4.1.3 Glycan microarrays applications:

The glycan microarray technology has several applications in diverse scientific areas [Palma, *et al*, 2015]. Here, we will highlight two microarrays objectives in this thesis.

Bacteroidetes species have PULs systems for the degradation of polysaccharides that host cells cannot hydrolyse. The produced CBM module from PUL BT0996 has Rhamnogalacturonan II as a validation substrate for culture growth [Martens, *et al*, 2011]. The initial approach was mostly to design a microarray set of pectins, hemicelluloses and yeast polysaccharides to identify the possible ligands. Simultaneously, the *N.oculata* fractions that were under characterization, were printed in this set to decode the presence of glycosidic linkages. This was performed by using characterized proteins that bind to specific carbohydrates and glycosidic linkages. We probed also the CBM module from PUL BT0996 to see if it has the ability to bind to different types of glycans, other than pectins (speculated to be their ligand, since it is the validation substrate of the PUL of this CBM).

After the initial screening in the manual arrayed microarray, a second glycan microarray was used to identify the specific epitope recognized by our CBM. Here, we used beside a greater diversity of glycans in the set, glycans with defined and known structure. In addition, we attempted to identify the ligands for the Bacova03100_A module in this array, since we had no validation substrate for the PUL of *B.ovatus* CBM module.

Here, our main objective was to determine the CBM-glycan interaction. The identification of the recognized epitopes will give some indication of the evolution process of these modules as well the elucidation of the polysaccharide degradation mechanisms by the PULs systems from the two Bacteroidetes species used in our studies.

4.2 Glycan microarray assay method:

4.2.1 Description of array assay surface used in this thesis:

In this thesis, the immobilization of the carbohydrates onto the nitrocellulose matrix coated glass slide was non-covalent and will be explained in detail.

Nitrocellulose matrix is, in these days, the most used membrane for the non-covalent immobilization and is based on the hydrophobic interaction between the glycans and the matrix. As described in section 4.1.2, low weight oligosaccharides (or disaccharides) must be derivatized as neoglycolipids to be printed [Liu, et al, 2007]. Here, each glass-slide is called the set, which represents the total number of glycans that are tested. Each slide contains a varied number of nitrocellulose membranes called pads, that are shown in figure 4.2.

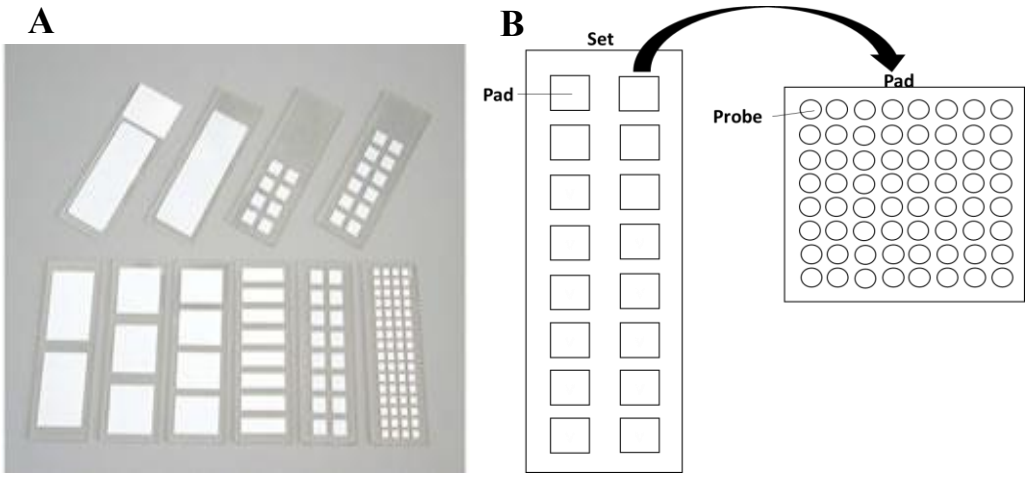


Figure 4.2- A-Example of nitrocellulose matrix coated glass slides. A glass slide can have one or diverse pads. B- Illustration of the composition of the sets.

The glycans are printed in the pads. Each printed glycan forms a spot and is called probe. A pad usually contains dozens of printed glycans, which normally are printed at two different concentrations for quantitation.

4.2.2 Choices of glycan libraries:

For the specificity analysis of the two produced CBM modules, two different glycan microarrays were constructed, one manually printed and other robotically printing. Using a robot array for the printing has the advantage of using small volumes of the probes. This means that more glycans can be printed and then simultaneously tested.

4.2.2.1 Manual assay construction:

In our laboratory, when using the manual printing, the glycans are immobilized in 1 or 2 pad slides. In this study we used the 2 pad slides.

The probes printed in this array were composed of polysaccharides from plants (such as pectins and hemicelluloses), from yeast and from microalgae, specifically from *N.oculata* cell walls. The pectins and hemicelluloses used have their main composition and structure determined, while the *N.oculata* polysaccharides, despite the determination of its major components by mass spectrometry, has the linkages between monomers to be completely assigned. The microalgae glycans were kindly provided from Universidade de Aveiro. For decoding the linkages, validated proteins were used and will be described in the next section. At the same time, we tested our Bt0996_C module to see if they could also interact with others glycans types that may be ingested and part of our diet. Hence, if the microalgae polysaccharides have the same linkages between the monomers that are found in the other polysaccharides used, there will be a cross-validation of both glycans types. On the other hand, if Bt0996_C module only recognizes these polysaccharides instead the characterized glycans, this would mean that the module interacts with other types of glycans (with different monosaccharides and glycosidic linkages) that are not present in the polysaccharides with known/determined structures that we used in the microarray.

The probe name, the set position and the predominant oligosaccharide sequence/monosaccharide composition of the glycans used are described in the following table 4.2.

Table 4.2- List of all glycans probes used in the binding charts and in the matrix (heat-map), position and the predominant sequence/ monosaccharide composition.

Probe	Set	Predominant oligosaccharide sequence/ Monosaccharide composition
Et50	1	Under characterization
Et85	2	Under characterization
Et85-1	3	Under characterization
Et85-2	4	Under characterization
Et85-3	5	Under characterization
EtSN	6	Under characterization
Pullulan	7	α-(1-4)-matrioses
α-Mannan	8	α-(1-3)-mannoses
Glucomannan	9	β-(1-4)-linked mannose and glucose
Galactomannan	10	β-(1-4)-linked mannose and glucose backbone α-(1-6)-galactose branches
Rhamnogalacturonan I	11	Mixed-linked α-(1-2;1-4) rhamnose and galacturonic acid backbone
Galacturonate LM	12	α 1-4 galacturonic acid backbone
Galacturonate HM	13	α 1-4 galacturonic acid backbone
PGA apple	14	α 1-4 galacturonic acid backbone
PGA citrus	15	α 1-4 galacturonic acid backbone
Oat β-glucan	16	Mixed-linked β-(1-3;1-4) glucoses

The Et fractions are from microalgae *N.oculata* cell walls, using a separation process done by ethanol precipitation.

As stated before, two plants glycan types were immobilized: Hemicelluloses and pectins. Hemicelluloses used in this manual microarray were Glucomannan, Galactomannan and Oat β-glucan. Pectins used were Rhamnogalacturonan I from soy bean, Galacturonate LM, Galacturonate HM, PGA from citrus and apple. In addition, two yeast polysaccharides, pullulan and α-Mannan were immobilized.

Each glycan probe was printed in duplicate at two concentrations: 0.1 and 0.5 mg/ml. A marker for the quality control, Cy3, was used for the glycan spots visualization.

4.2.2.2 Robotic assay construction:

Glycan immobilization using a robot has the advantage that a larger number of glycans can be printed and more proteins can be tested simultaneously. The slides were printed in the Imperial college by Dra. Angelina Palma, in collaboration with the group of Prof. Ten Feizi at the Glycosciences Laboratory, Imperial College London. Here, the glycans were immobilized into 16 pads slides. This set consisted of newly characterized pectins, acidic glycans and mammalian glycans (table 4.3). Furthermore, pectins and *N.oculata* polysaccharides previously tested in the manual array were also included, for cross-validation.

Table 4.3- List of all glycans probes used in the binding charts and in the matrix (heat-map), position and the predominant sequence/ monosaccharide composition.

Probe	P[a]	P[b]	Predominant oligosaccharide sequence/ Monosaccharide composition
PGA (Citrus)	1	17	α 1-4 galacturonic acid backbone
Galacturonate LM (Apple)	2	18	α 1-4 galacturonic acid backbone
Galacturonate LM (Citrus)	3	19	α 1-4 galacturonic acid backbone
Pectin Galactan (Lupin)	4	20	β1-4 Galactose (Gal: Ara: Rha: Xyl: GalUA = 77: 14: 3: 0.6: 5.4
Pectin Galactan (Potato)	5	21	β1-4 Galactose (Gal: Ara: Rha: GalUA = 78: 9: 4: 9)
Galactan (Lupin)	6	22	β1-4 Galactose (Gal:Ara:Rha:Xyl:other sugars = 82 : 5.8 : 5.1 : 1.4 : 5.7, Galacturonic acid 14.6%
Rhamnogalacturonan I (soy bean)	7	23	Mixed-linked α-(1-2;1-4) rhamnose and galacturonic acid backbone

50WSnFl-S2 (S. nigra)	8	25	Ara (28%), Rha (4.4%), Xyl (1.3%), Man (1%), Gal (19.2%), Glc (2%), GlcA (0.4%), GalA (42.3%), 4-O-Me-GlcA (1.4%)
100WSnFl-S2 (S. nigra)	9	26	Ara (18.2%), Rha (16.8%), Xyl (3.4%), Man (0.5%), Gal (17.8%), Glc (2.8%), GlcA (0.3%), GalA (40.5%), 4-O-Me-GlcA (0.9%)
50WSnFl-S2-EI (S. nigra)	10	27	Ara (29.5%), Rha (14.3%), Fuc (0.4%), Xyl (1.8%), Man (2.0%), Gal (25.5%), Glc (3.6%), GlcA (2.3%), GalA (17.9%), 4-O-Me-GlcA (2.7%)
SnFl50-S2 (S. nigra)	11	32	Ara (19.4%), Rha (5.3%), Xyl (0.7%), Man (1.1%), Gal (22.9%), Glc (2.8%), GlcA (2.1%), GalA (44.7%), 4-O-Me-GlcA (1%)
IOI-WAc (I. obliquus)	12	29	Under characterization
IOI-WAc (I. obliquus)	13	30	Under characterization
BP-II (B. petersianum)	14	33	Ara (5.1%), Rha (8.2%), Fuc (0.5%), 2-Me-Fuc (trace), Xyl (6.3%), 2-Me-Xyl (trace), Man (0.7%), Gal (8.3%), Glc (4.4%), GlcA (1.3%), GalA (65.1%)
GOA1 (G. oppositifolius)	15	34	Ara (26.4%), Rha (4.2%), Xyl (3.9%), Man (4.3%), Gal (42.9%), Glc (3.5%), GalA (12.1%), 4-O-Me-GlcA (2.9%)
GOA2 (G. oppositifolius)	16	35	Ara (5.5%), Rha (10.3%), Fuc (1.3%), Xyl (0.5%), Man (0.6%), Gal (9.7%), Glc (3.3%), GalA (68.3%), 4-O-Me-GlcA (0.4%)
Vk100-Fr.I (V. kotschyana)	17	36	Ara (2%), Rha (1%), Fru (83%), Gal (2%), Glc (3%), GalA (1%)
Ctw-A1 (C. tinctorium)	18	37	Ara (16.3%), Rha (17.9%), Man (1.8%), Gal (45.(8%), Glc (4%), GlcA (8.8%), GalA (5.8%), Fru 4.9%)
Oc50A1.IA (O. celtidifolia)	19	38	Ara (38.9%), Rha (4.2%), Man (5.8%), Gal (30.9%), Glc (5.4%), GlcA (trace), GalA (11.5%), 4-O-Me-GlcA (3.3%)
LPS3 (T. cordata)	20	39	Under characterization
LCC (C. cordifolia)	21	60	Under characterization
CC1P1 (C. cordifolia)	22	31	Ara (trace), Rha (32%), Gal (31%), Glc (2%), GalA (35%)
CC1 (C. cordifolia)	23	40	Ara (3.7%), Rha (22.1%), Gal (20.2%), Glc (0.5%), GalA (29.6%), 2-O-Me-Gal (6.5%), 4-O-Me-GlcA (17.4%)
CC2 (C. cordifolia)	24	41	Ara (37.2%), Rha (8.5%), Gal (31.3%), Glc (1.1%), GalA (11.5%), GlcA (3.4%), 2-O-Me-Gal (0.4%), 4-O-Me-GlcA (6.6%)
CC3 (C. cordifolia)	25	42	Ara (3.0%), Rha (22.8%), Gal (17.3%), Glc (1%), GalA (32.8%), 4-O-Me-GlcA (17.8%), 2-O-Me-Gal (5.3%)
PBS100-II (P. biglobosa)	26	43	Ara (21.2%), Rha (7.3%), Xyl (0.2%), Gal (18%), Glc (6.1%), GalA (30.1%), GlcA (10.5%), 4-O-Me-GlcA (1.3%)
CSA (Sigma C8529)	27	55	mixed-linked-β-(1-3,1-4) glucuronic acid and N-acetyl-galactosamine-4-sulfate
CSB (Sigma C2413)	28	56	β1-3 L-iduronic acid and N-acetyl-galactosamine-4-sulfate
CSC (Sigma C4384)	29	57	mixed-linked-β-(1-3,1-4) glucuronic acid and N-acetyl-galactosamine-6-sulfate
HA (Sigma H7630)	30	58	mixed-linked-β-(1-3,1-4) glucuronic acid and N-acetyl-galactosamine
Fraction Et50 (N. Oculata)	31	24	Under characterization
Fraction Et85 (N. oculata)	32	44	Under characterization
Fraction Et85-1 (N. oculata)	33	45	Under characterization
Fraction Et85-2 (N. oculata)	34	46	Under characterization
Fraction Et85-3 (N. oculata)	35	47	Under characterization
Fraction EtSU (N. oculata)	36	48	Under characterization
Galactoglycolipid (N. oculata)	37	59	Under characterization
Xylan X3 (Plum)	38	49	Rha (8%), Fuc (0%), Ara (13%), Xyl (69%), Man (2%), Gal (3%), Glc (1%), Ur Ac (3%)

Xylan X4 (Plum)	39	50	Rha (0%), Fuc (0%), Ara (11%), Xyl (59%), Man (0%), Gal (2%), Glc (4%), Ur Ac (25%)
Xylan X5 (Plum)	40	51	Rha (2%), Fuc (2%), Ara (10%), Xyl (63%), Man (0%), Gal (5%), Glc (5%), Ur Ac (13%)
Xyloglucan XG3 (Plum)	41	52	Rha (2%), Fuc (6%), Ara (5%), Xyl (44%), Man (4%), Gal (12%), Glc (25%), Ur Ac (2%)
Xyloglucan XG4 (Plum)	42	53	Rha (2%), Fuc (5%), Ara (3%), Xyl (36%), Man (7%), Gal (11%), Glc (29%), Ur Ac (8%)
Xyloglucan XG5 (Plum)	43	54	Rha (4%), Fuc (5%), Ara (5%), Xyl (35%), Man (5%), Gal (11%), Glc (28%), Ur Ac (4%)

P^a- Position matching the graph/heatmap
P^b- Position matching the original microarray set

Proteins for microarray validation also included the mouse anti-His, the biotinylated antibody responsible for the detection of the primary antibody (rat antibody). It had the goal to exclude probes in our analysis that fluorescence signal would be observed.

4.2.3 Glycan microarrays binding assay:

4.2.3.1 Theory:

Prior to the start of the assay, the slides are scanned for Cy3 fluorophore at 532 nm for glycan spots visualisation.

At the beginning, each pad of the slide is dry and must be wetted with water, or otherwise the matrixes are too hydrophobic and the proteins will not interact with the glycan probes. Thereafter, a blocking solution is added onto the pads and incubated usually for 1 hour. This step is crucial to avoid non-specific interactions by any particle that may be present in solution.

The blocking solution in our laboratory is composed of proteins such BSA and/or Casein, that are previously made in HEPES saline buffer (HBS). Using the correct blocking solution composition is a “trial and error” procedure. It influences the result when the slides are scanned for Alexa Fluor-647. The background can be too high for performing a correct quantification.

After removal of the blocking solution, the binding step proceeds. Here, 3 types of procedures are performed based on the protein type (rat antibody, CBM module or lectins) that are probed in the microarray slides (figure 4.3).

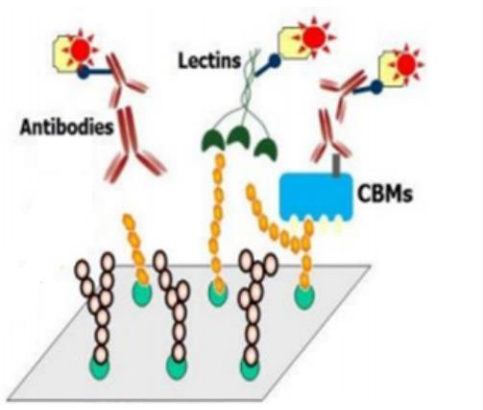


Figure 4.3- Illustration of glycans binding step by different proteins: antibodies, lectins and CBM modules.

For rat antibodies sample, it is usually incubated for 1 hour onto the pad. However, for the detection of the glycan-antibody interaction, a second antibody (mouse antibody) is afterward incubated for 1 hour that is against the rat antibodies. Since the secondary antibody is biotinylated, streptavidin conjugated with Alexa Fluor-647 is added and the slide is scanned at 647 nm, which emits red spots corresponding to glycan-sample interactions.

For the CBMs sample, the procedure is identical, with the addition of a rat antibody, called primary antibody, which is against the His-tag of the recombinant CBM module. The remaining procedure is described above.

Lectins, in turn, have the easiest binding step. They are already biotinylated and don´t require the use of any antibody to be detected. As we did to the other two protein types, streptavidin is added later.

4.2.3.2 Experimental:

All solutions in this thesis were prepared with the corresponding blocking solution at the desired concentrations (index table 4)

4.2.3.2.1 Manual array (slide of 2 pads):

The pads were scanned for Cy3 at 532 nm (Molecular Devices) and wetted with 500 µl of Mili-Q water. The water was removed and 500 µl of corresponding blocking solutions were added and incubated for 1 hour at room temperature.

After removing the blocking solutions, 500 µl of binding solutions were added and incubated for 1 hour and 30 minutes at room temperature. To shorten the assay time, pre-complexes were previously made, which consisted on preparing solutions with the protein and the antibodies already mixed and incubated for 15 min before being applied onto the pads, and thus, reducing the number of incubations from three to one.

After the binding step, 500 µl of streptavidin (0.5 µg/ml) were added and incubated for 45 min. The detection of the signal was performed by scanning the pads for Alexa Fluor-647 at 647 nm, using GenePicPro7 software.

Each time an incubation was performed, the pads were washed four times with HBS. At the end of the assay, the slides were washed four times with HBS, then two times with Mili-Q water, followed by 1 minute of centrifugation. The slides were then placed in the dark for 5 min to dry, before being scanned.

4.2.3.2.2 Robotic array (slide of 16 pads):

The procedure was like the used for the manual array, except that in the 16 pads slides, the pad size is smaller and requires a lower solution volume (90 µl) to be completely wet.

4.2.4 Glycan microarrays analysis:

GenePixPro 7 software allows the user to choose the laser wavelength (mentioned in the experimental section), the laser power (%) and the corresponding emission filter. The choice of laser power for the acquisition of the images impacts the quantitation result. The interaction spot should not saturate and, on the other hand, background intensity should be as low as possible.

The data obtained (the images) is processed into three steps: first, the construction of the grid for each pad; second, the quantitation of fluorescence intensity; third, the presentation of the results.

The grid construction was performed using the Cy3 image scan (while the assay was occurring) and adjusted to the Alexa Fluor-647 image scan. The grid is essential to indicate to the program which are the zones for quantifying the fluorescence intensities. Each spot is then subtracted with the fluorescence from outside of the spot (background). The result will be a numeric value, saved as a GenePix results file (gpr).

A custom software from the Glycosciences Laboratory, from the Imperial College London, was developed to process this numeric value. It integrates 2 input programs (Microarray Database and Piezo array) and 1 output program. Microarray Database has the information of each set (name of the probes and predominant oligosaccharide sequence/monosaccharide compositions). When a new set is used for the first time, the information is entered in the Microarray Database for set identification. Piezo array allows the input of all the experimental data such as the scanned

images, grids, gpr files, proteins batches and concentrations, probe set used, information of the blocking solution used and the realization date. The integration of the results into these input programs is retrieved from the Display array (output). This program allows the presentation of the processed results in form of tables, charts and matrices.

4.3 Results and Discussion:

4.3.1 Manual array results:

The processed data of manual array in this thesis are presented in form of graphics (figures 4.4 and 4.5) and a heat-map (figure 4.6). Proteins chosen for validation recognize a large diversity of glycans, immobilized in our array. These validation proteins consisted of 2 antibodies (α 1,3/1,4-β-glucan and LM21), 2 biotinylated lectins (Concanavalin A and *Ricinus Communis Agglutinin* I) and 1 CBM module from *Cellvibrio Mixtus*.

The validation proteins are going to be described first for the quality control, followed by the interaction analysis of Bt0996_C module, and finally, the final observations of the manual microarray, showing the heatmap.

4.3.1.1 Proteins for quality control analysis:

In this section, each graph for a validation protein is shown in figure 4.4. Each validation protein is going to be described individually for a better interpretation.

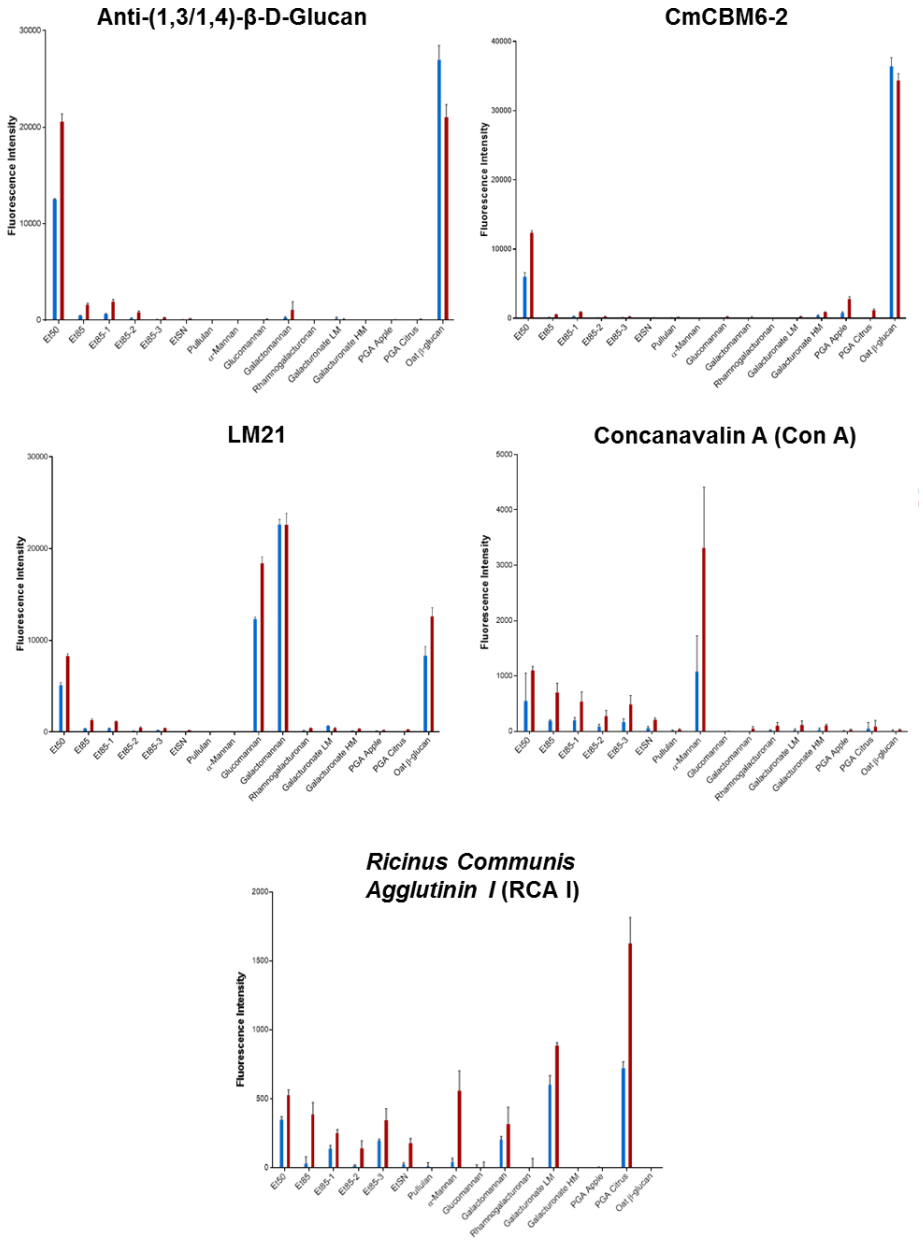


Figure 4.4- Glycan microarray data analysis of proteins for que quality control of the microarray set. Protein names are written on top of each graph and each recognize different epitopes and have different specificities (index table 3). Glycans sequence information of the probes included in the microarray are in Table 4.2. The interacting signals are the fluorescence intensities of duplicate spots of probe with error bars. Each probe was printed at two concentrations: 0.1 and 0.5 mg (dry weight). Quatified fluorencense intensity is plotted on the y-axis. Glycan probes are plotted on the x-axis.

Anti-1(1,3/1,4)-β-D-Glucan is an antibody that is high specific for mixed linked 1,3/1,4 glucoses, as described in the literature [Burton, *et al*, 2009]. Oat β-glucan is the validation glycan used that contains the mentioned linkage. As expected, the antibody recognized strongly the glycan, confirming thus its specificity (figure 4.5)**.** Fluorescence signal was also strongly observed for Et50, a polysaccharide fraction from microalgae *N.oculata* that is currently under characterization. Due to the strong signal, the glycan may have in its composition mixed linked β 1,3/1,4 glucoses.

The **CmCBM6-2** module has the same glycan specificity [Henshaw, *et al*, 2004] as the α 1,3/1,4-β-glucan antibody and the fluorescence signals observed were the same as for the antibody (figure 4.6). The use of a characterized CBM was a quality control element that ensured our protein type (CBM module) could interact perfectly with the probes.

LM 21 is a characterized antibody that binds to β1-4 linked mannose and glucose [Marcus, *et al*, 2010]. Seeing the figure 4.7, interactions were observed for glucomannan and galactomannan, both with the same backbone, that is the epitope of this antibody. However, moderate intensities signals were observed for Oat β-glucan and Et50. This may since at the binding step, the assay with the antibody α 1,3/1,4-β-glucan was performed in the same slide and may have contaminated the other pad, which recognizing thus these 2 glycans.

Con A is a lectin that has preference for β-1-4 linked mannoses in the terminus of the glycans [Wang, *et al*, 2014]. Strong signal was observed for α-Mannan (figure 4.8), that has a similar epitope which is β 1-3 linked mannoses. Furthermore, it also recognized Et85-3 fraction. Although this glycan is under characterization, this glycan could have β- linked mannoses in the terminals.

RCA 120 prefers terminal β- linked galactoses [Wang, *et al*, 2011]. Fluorescence intensities were observed for PGA citrus and Galacturonate LM (figure 4.9), both with the main component being α 1-4 linked galacturonic acids.
Galacturonic acid is the galactose oxide form. However, due to the low values of fluorescence intensity (around 1000-2000), it may suggest RCA I was interacting either with a small number of branches that have terminal galactoses, or, with any remaining contaminant in these glycan samples, since both these glycans came from natural sources.
RCA 120 also interacted with galactomannan, but didn´t bind to glucomannan, because it lacked branches with galactose and mannose linked by a β 1-6 linkage. In addition, it was the only protein that had capacity to bind to EtSN fraction (although a weak interaction).
Although a signal has been observed for α-Mannan antibody, it may due to the Con A having passed to the pad.

Validated the glycan microarray, the analysis of His-Bt0996_C module is going to be described and its possible epitope speculated.

4.3.1.2 Bt0996_C module analysis:

B.thetaiotaomicron doesn´t have PULs with specific mechanisms for the degradation of hemicelluloses. That said, it was expected that our module wouldn't interact with Glucomannan, Galactomannan or Oat β-glucan.

Our CBM module strongly recognized Low Metilated Galacturonate and PGA Citrus (figure 4.5), both with α 1-4 linked galacturonic acids, as referred above. The reason for not recognizing Galacturonate High Methylated and PGA apple is probably because the rate of methylation in these polysaccharides is higher, blocking the interaction. Same is applied to the lectin RCA I, which is described above.

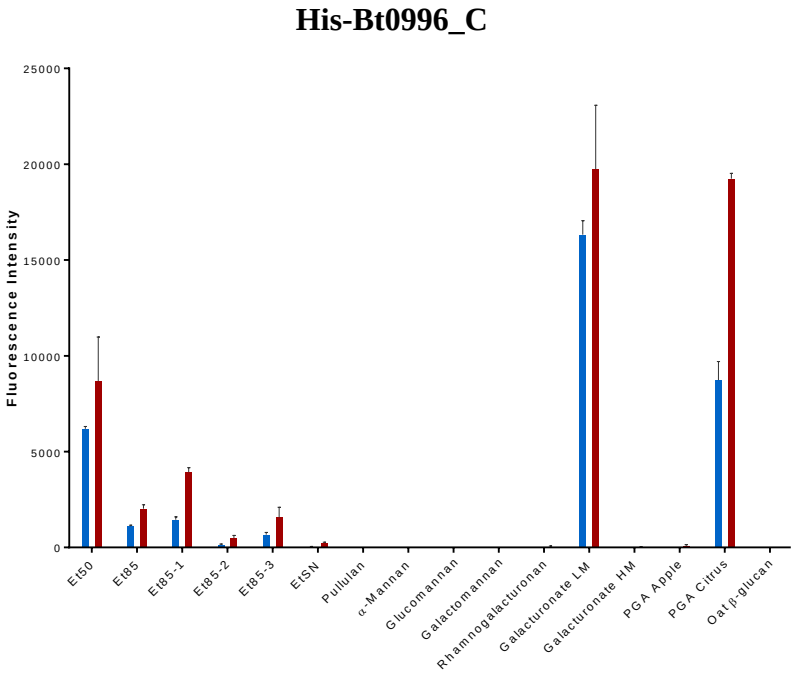


Figure 4.5- Glycan microarray data analysis for our His-Bt0996_C module in study. Module name is written on top of the graph. Glycans sequence information of the probes included in the microarray are in Table 4.2). The interacting signals are the fluorescence intensities of duplicate spots of probes with error bars. Each probe was printed at two concentrations: 0.1 (blue bars) and 0.5 mg (red bars) at dry weight. Quantified fluorescence intensity is plotted on the y-axis. Glucan probes are plotted on the x-axis.

Due to the observed strong signals with Galacturonate LM and PGA citrus, it seems that our CBM module interacts with α 1-4 linked galacturonic acids. However, it is possible that branches are present in Galacturonate LM and PGA and that the CBM module recognizes the branch instead of the glycan backbone.

The CBM module recognized more weakly the *N.oculata* polysaccharides fractions. To be sure that the interactions were specific, this result will be discussed later, together with the robotic microarray results. There, we used the protein with two different His-tag positions, N- and C-terminal, to compare the interactions observed.

4.3.1.3 Comparison of the proteins signal spots

In this section, the comparison between proteins specificities is shown in the heatmap (figure 4.6).

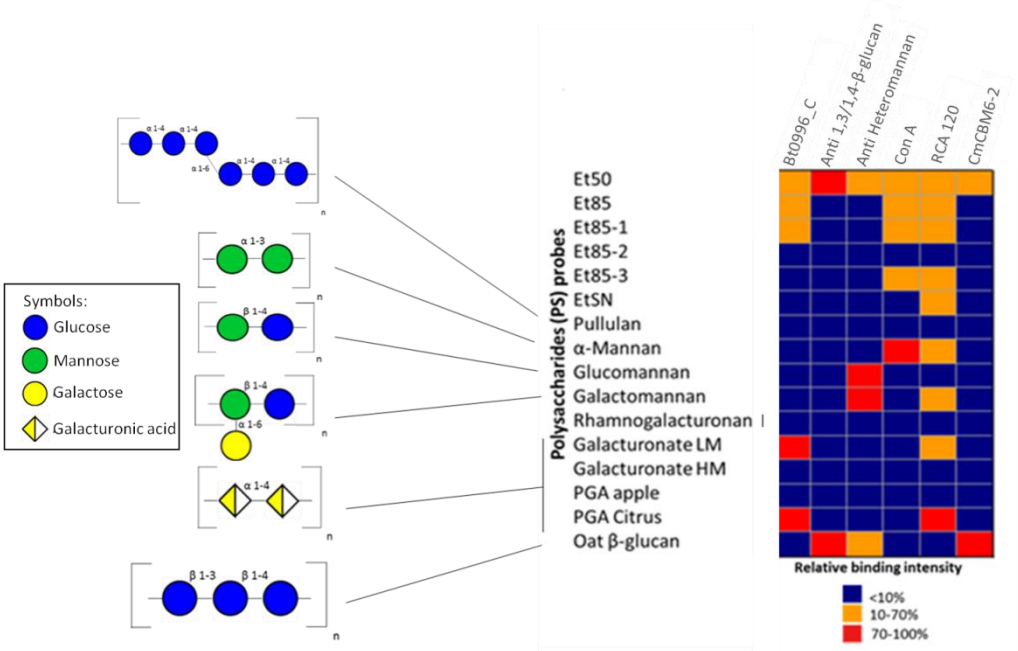


Figure 4.6- Heat-map analysis of the relative binding intensities calculated as the percentage of the fluorescence signal intensity given by the probe most strongly bound by each protein (normalized as 100%). Blue spots- interactions below 10%; White spots- interactions between 10-30%; Orange spots- Interactions between 30-70%, Red spots- Interactions above 70%.

As described previously, each characterized protein bound to the respective specific glycans to which it has specificity, and thus, the large diversity of binding interactions is shown in the heat-map. The unexpected fluorescence intensities, for LM21 to Oat β-glucan and for RCA I to α-Mannan, were due to contamination of the pad by neighbouring proteins.

Pullulan, a yeast polysaccharide, was used as negative control. As expected, no interaction signals were observed by the characterized proteins used in this assay. Neither Bt0996_C module recognized this glycan.

Bt08996_C module bound strongly only to PGA Citrus and Galacturonate LM. Bt08996_C and RCA I have an identical binding intensity pattern for the characterized glycans, except for galactomannan, which suggests that the epitope recognition may be different. An interesting result was that Bt0996_C didn´t recognize Rhamnogalacturonan I, while the Rhamnogalacturonan II is the validation substrate for *B.thetaiotaomicron* growth. A possible explanation will be discussed in section 4.3.2.
To fully identify and confirm the epitope recognized by the Bt0996_C CBM module, a 16-pad glycan microarray was robotically printed with other pectins and characterized oligosaccharides. This time, the number of probes to be tested had a larger diversity.

4.3.2 Robotic array analysis:

4.3.2.1 Proteins for quality control analysis:

The graphs of the validation proteins (A) and the proteins under characterization of glycan binding specificity (B) are shown in figure 4.7.

The two antibodies (primary and secondary) from the detection system were tested (index x) since it had shown interactions with carrageenans (kappa, Iota and lambda) and *A.cepa* pectin. The fluorescence intensities for these probes didn´t significaty change with incubation of other proteins, thus these probes were excluded from our analysis.

Here, as in the manual printed microarray, we are going to first describe the quality control that was carried out using the 2 antibodies (Anti-(1,3/1,4)-β-D-Glucan and Anti-1,4-β-Galactan) and 1 biotinylated lectin (RCA I).

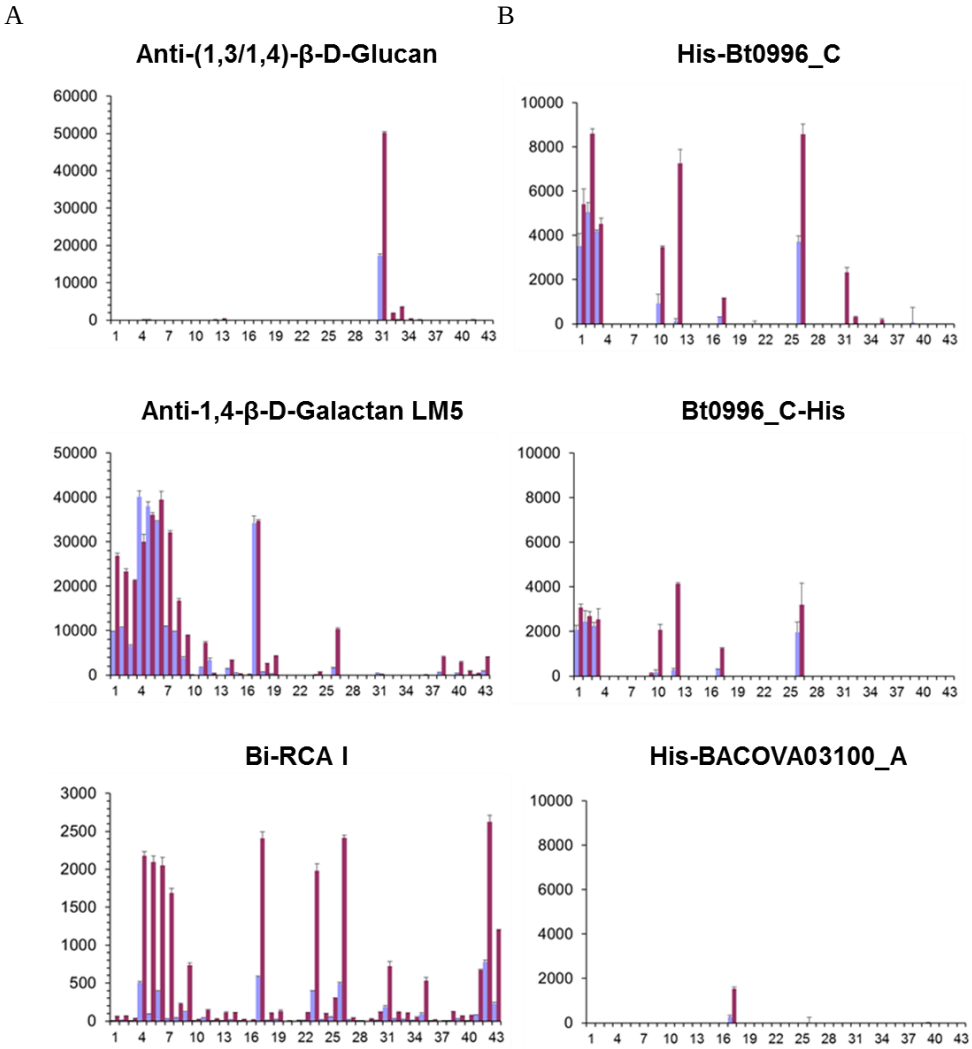


Figure 4.7- Glycan microarray data analysis of proteins for que quality control of the microarray set (A) and of proteins for characterization studies (B) . Protein names are written on top of each graph and each recognize different epitopes and have different specificities (index x). Glycans sequence information of the probes included in the microarray are in Table 4.2. The interacting signals are the fluorescence intensities of duplicate spots of probe with error bars. Each probe was printed at two concentrations: 0.1 and 0.5 mg (dry weight). Quatified fluorencense intensity is plotted on the y-axis. Glycan probes are plotted on the x-axis.

Anti-(1,3/1,4)-β-D-Glucan as stated before, it has the specificity of mixed linked β-1,3/1,4 glucoses [Burton, *et al*, 2009]. In this robot microarray, a strong and unique interaction was observed with a *N.oculata* fraction, Et50 (figure 4.7; position 31). This result is in agreement with manual array result, thus, confirming the possibility of the ET50 fraction having as main component mixed linked β-1,3/1,4 glucoses.

Anti-1,4-β-D-galactan is an antibody that recognizes galactoses linked with β-1,4 linkage [Jones, *et al*, 1997]. As shown in the graph, it bound to a large diversity of glycans. Strong fluorescence intensities (figure 4.7; positions 4-7 and 17) were observed for Galactan from Lupin and potato, Rhamnogalacturonan I from Soy bean and a pectin from *V. kotschayana*. Except for *V. kotschayana* pectin, each of the glycans mentioned have β-1,4 linked galactoses as their main oligosaccharide sequence , and thus this result was expected. We suspect that the *V.kotschayana* signal is due to a contaminant containing β-1-4 linkages and thus that the interaction is non-specific.

Moderate signals (figure 4.7; positions 1-3 and 8) were observed for PGA citrus, Galacturonate LM (from citrus and apple) and a *S. nigra* fraction, 50WSnFI-S2. Although the principal linkage of these pectins is α 1-4 galacturonic acid, they have a moderate content of galactose monomers, and may have its specific glycosyl linkage.

This antibody also bound to, although weakly (figure 4.7; positions 9, 11, 19, 26, 38 and 43) to two *S. nigra* fractions, 100WSnFI-S2 and SnFI50-S2, to the *O. celtidifolia* pectin, to the *P. biglobosa* pectin, to Xylan X3 and to Xyloglucan XG. This could have been due to the presence of minor percentage β-1,4-linked galactoses in these glycans, since the galactose percentage is moderate (table 4.3; values around 20%).

The biotinylated **RCA I,** as already described, binds to terminal galactoses [Wang, *et al*, 2011]. It showed interactions (figure 4.7; positions 5-8, 17, 23, 42 and 43) with Galactans from lupin and potato, Rhamnogalacturonan I from soy bean, *V. kotschyana* pectin, a *C. cordifolia* fraction, CC1, *P. biglobosa* pectin, Xyloglucan XG4 and Xyloglucan XG5. These glycans (except V. kotschyana pectin) have a moderate galactose content (table 4.3; values around 10-20), thus explaining the more moderate signal observed.

There were observed weak interactions (figure 4.7; positions 9, 25, 31, 35 and 41) for another *S. nigra* fraction, 100WSnFI-S2, another *C. cordigolia* fraction, CC3, few *N. oculata* fractions, Et50 and Et85-3, and *Xyloglucan XG3*. The S. nigra fraction, 100WSnFI-S2 has a moderate galactose content (table 4.3; value of 17%), that must be present at the terminals of the pectin structure to give binding signal.

The others glycans (except *N. oculata* fractions, which were under characterization) have a relatively low galactose content (table 4.3- values below 12%), and explaining the observed weak spots. However, this signal interaction could also be from any contaminant that remained in the glycan sample.

In sum, the validation proteins showed glycans specificities that is according to literature. Therefore, we will now discuss in more detailed manner the interactions observed with our CBM modules.

4.3.2.2 CBM modules analysis:

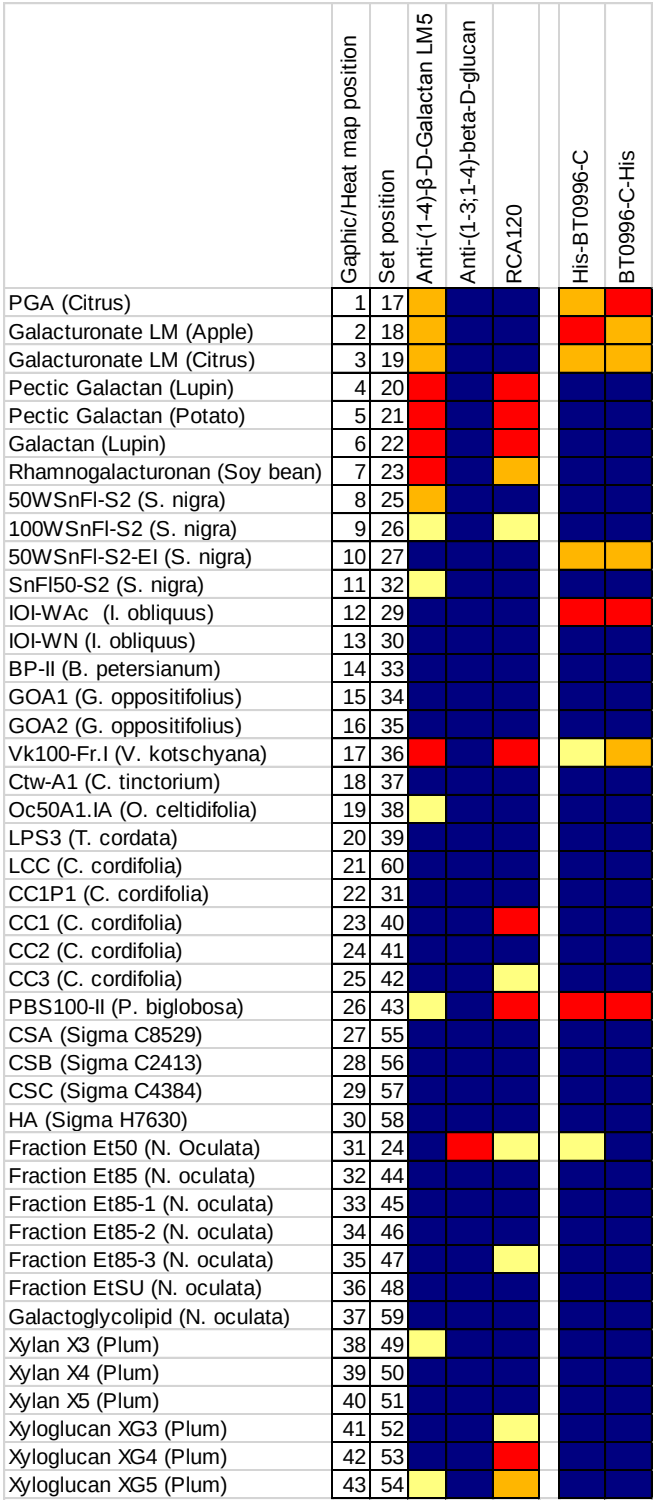
In a first analysis, the two CBM module constructs, **Bt0996_C** with N- and C-terminal His-tag, showed identical binding patterns. The exception was for the fraction of *N.oculata*, Et50. In the manual arrayed microarray, we thought that this CBM module recognized this fraction. However, in the robotic microarray, Bt0996_C-His-tag module didn´t showed again this interaction, meaning that the position of the His-tag influenced the interaction of the glycan fraction, being thus non-specific.

Interactions were shown (figure 4,7; positions 1-3; 10, 13, 17 and 26) with PGA citrus and Galacturonate LM (citrus and apple), as observed in the manual printed array. In adition, it recognized four other pectins: 50WSnFI-S2-EI fraction from *S. nigra*, IOI-WAc fraction from *I. obliquus*, Vk100-Fr.I fraction from *V. kotschayana* and PBS100-II fraction from *P. biglobosa*. The *I. obliquus* fraction, IOI-WAc, is under characterization, whereas the *V. kotschayana* fraction, Vk100-Fr.I, is known to have a high frutose content (table 4.3; value of 83%). Here, we suspected that the interaction was non-specific, since for other glycans, there was no fructose present in the sample. Both *P. biglobosa* and *S. nigra* fractions have a moderate content of galacturonic acid (table 4.3; values around 18-30%). An interesting result is that our CBM module recognize only one of the *S. nigra* fractions, 50WSnFI-S2-EI. Each *S. nigra* fraction, as stated, has a moderate content of galacturonic acid (values around 18-30%). Significative changes between fractions are observed for the arabinose and rhamnose content. The 50WSnFI-S2-EI fraction has a high rhamnose content, the highest arabinose and the lowest galactoronic acid contents. We speculate that the interation by this CBM module will be with an oligosaccharide composed of galacturonic acid and rhamnose, not present in rhamnogalacturonan I.

With the His-BACOVA03100_A module, despite the large diversity of glycans printed in the array, no interactions were observed. Other glycan microarray sets are under evaluation.

4.3.2.3 Comparison of the proteins signal spots:

The analysis of the relative binding intensities for each protein (excluding His-BACOVA03100_A) is compared in the figure 4.8.



	Gaphic/Heat map position	Set position	Anti-(1-4)-β-D-Galactan LM5	Anti-(1-3;1-4)-beta-D-glucan	RCA120	His-BT0996-C	BT0996-C-His
PGA (Citrus)	1	17					
Galacturonate LM (Apple)	2	18					
Galacturonate LM (Citrus)	3	19					
Pectic Galactan (Lupin)	4	20					
Pectic Galactan (Potato)	5	21					
Galactan (Lupin)	6	22					
Rhamnogalacturonan (Soy bean)	7	23					
50WSnFl-S2 (S. nigra)	8	25					
100WSnFl-S2 (S. nigra)	9	26					
50WSnFl-S2-EI (S. nigra)	10	27					
SnFl50-S2 (S. nigra)	11	32					
IOI-WAc (I. obliquus)	12	29					
IOI-WN (I. obliquus)	13	30					
BP-II (B. petersianum)	14	33					
GOA1 (G. oppositifolius)	15	34					
GOA2 (G. oppositifolius)	16	35					
Vk100-Fr.I (V. kotschyana)	17	36					
Ctw-A1 (C. tinctorium)	18	37					
Oc50A1.IA (O. celtidifolia)	19	38					
LPS3 (T. cordata)	20	39					
LCC (C. cordifolia)	21	60					
CC1P1 (C. cordifolia)	22	31					
CC1 (C. cordifolia)	23	40					
CC2 (C. cordifolia)	24	41					
CC3 (C. cordifolia)	25	42					
PBS100-II (P. biglobosa)	26	43					
CSA (Sigma C8529)	27	55					
CSB (Sigma C2413)	28	56					
CSC (Sigma C4384)	29	57					
HA (Sigma H7630)	30	58					
Fraction Et50 (N. Oculata)	31	24					
Fraction Et85 (N. oculata)	32	44					
Fraction Et85-1 (N. oculata)	33	45					
Fraction Et85-2 (N. oculata)	34	46					
Fraction Et85-3 (N. oculata)	35	47					
Fraction EtSU (N. oculata)	36	48					
Galactoglycolipid (N. oculata)	37	59					
Xylan X3 (Plum)	38	49					
Xylan X4 (Plum)	39	50					
Xylan X5 (Plum)	40	51					
Xyloglucan XG3 (Plum)	41	52					
Xyloglucan XG4 (Plum)	42	53					
Xyloglucan XG5 (Plum)	43	54					

Figure 4.8- Heat-map analysis of the relative binding intensities calculated as the percentage of the fluorescence signal intensity given by the probe most strongly bound by each protein (normalized as 100%). Blue spots- interactions below 10%; White spots- interactions between 10-30%; Orange spots- Interactions between 30-70%, Red spots- Interactions above 70%.

The heat-map showed high diversity of signal spots for each protein. Anti-(1-3,1-4)-β-Glucan showed an high specificity for fraction Et50 (from *N.oculata*). As discussed above, this glycan may have mixed-linked-β-(1-3,1-4) glucoses as main component. The restricted specificity of the antibody proved the quality control of the assay (as well as the other characterized proteins).
PGA citrus and Galacturonate LM from citrus and apple, although were recognized by Anti-(1-4)-β-D-Galactan and our module Bt0996_C, the binding patterns are different, and thus the recognition were by different epitopes.

As stated in the manual array and observed in this heat-map, Bt0996_C module didn´t bind to rhamnogalacturonan I. To understand the differences between RGI and RGII, we present the figure 4.9 that shows their schematic structure.

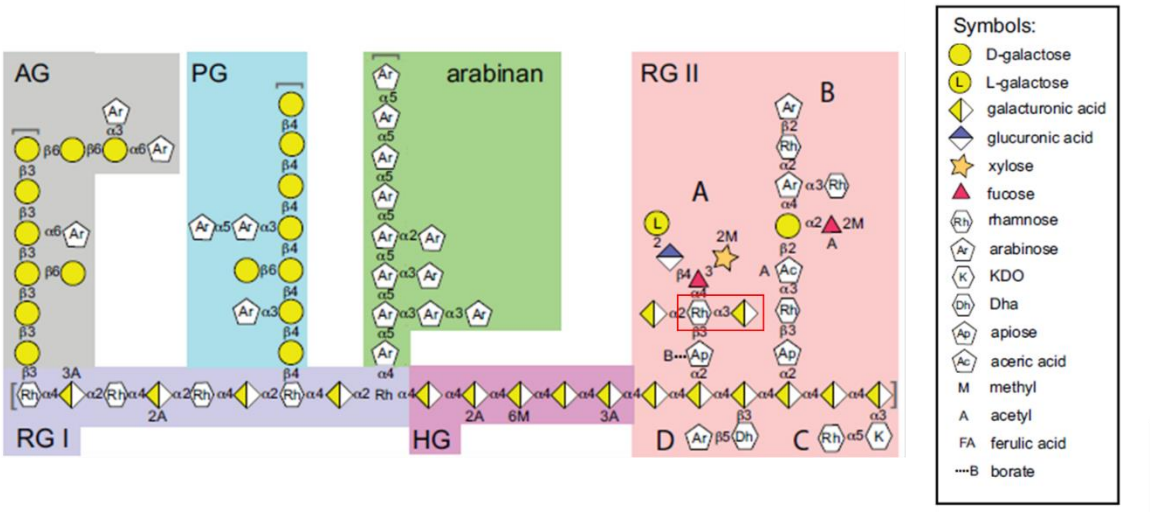


Figure 4.9- Schematic structure of RGI and RG II. RG I is associated with diverse pectins, such as arabinogalactan, pectin galactan and arabinan. Red rectangle in the symbolic structure of RG II is the interaction speculated by Bt0996_C module. Image adopted from article: Recognition and Degradation of Plant Cell Wall Polysaccharides by two Human Gut Symbionts.

The two structures are very different from each other. The backbone of RG I is composed of mixed linked α-(1-2,1-4) rhamnose and galacturonic acid while RG II is composed exclusively of α-(1-4) galacturonic acid. In addition, the side branches have several differences, as shown in the figure 4.9.

Two characterized proteins, Anti-(1-4)-β-D-Galactan and RCA I, can bind to RG I (figure 4.8). The Anti-(1-4)-β-D-Galactan has as recognition epitope the β 1-4 linked Galactose, and binds also to the pectin galactan (PG). The RCA I has the preference for β-(1-4) linked galactoses and binds to the PGs terminals. It is possible that RCA I may also interact, although weaker, with the terminal galactoses of the arabinogalactan.

The RG II don´t have these side branches. In fact, it has highly complex branches. Initially, we though that with the manual printed array that our Bt0996_C bound to α-(1-4) galacturonic acid (to the backbone). However, in the robotic array, the module didn´t recognize all the *S. nigra* fractions, that were essentially composed of α-1-4 galacturonic acid [Barsett, *et al*, 2012]. Here, we speculate that our CBM module binds to rhamnose and galacturonic acid linked by α-(1-3) linkage.

Despite the results obtained, further studies are required to decode the specificity of this Bt0996_C module.

4.4 Conclusions

The main objective of the glycan microarray analysis is to understand the diversity of glycan monosaccharides sequences and their recognition by different proteins, especially by our CBMs modules.

The manual printed microarray was effective to initially identify some possible ligands of the Bt0996_C module. However, we weren´t successful in fully discovering the specificity of the module in the robotically arrayed microarray.

A future perspective is to, with the collaboration of Dr. Harry Gilbert, produce the various fragments of RG II, using *B. thetaiotamicron* with mutated PULs [Ndeh, *et al*, 2017],thus printing a new set with these fragments. This will lead to discovery of fragment recognized by the module, and eventually, to its specificity assignment.

Chapter 5- Structural characterization of CBMs using X-ray crystallography

5.1 Introduction:

Discovering the protein 3D structure is essential for understanding its function. Nowadays, there are several powerful techniques for protein structure determination such as X-ray crystallography, nuclear magnetic resonance (NMR) or cryo-electron microscopy (Cryo-EM). Here, X-ray crystallography technique was used and thus will be explained in more detail.

The X-ray crystallography method involves protein crystallisation, X-ray diffraction, data processing and model building (illustration in figure 5.1) [Papageorgiou, *et al* 2014].

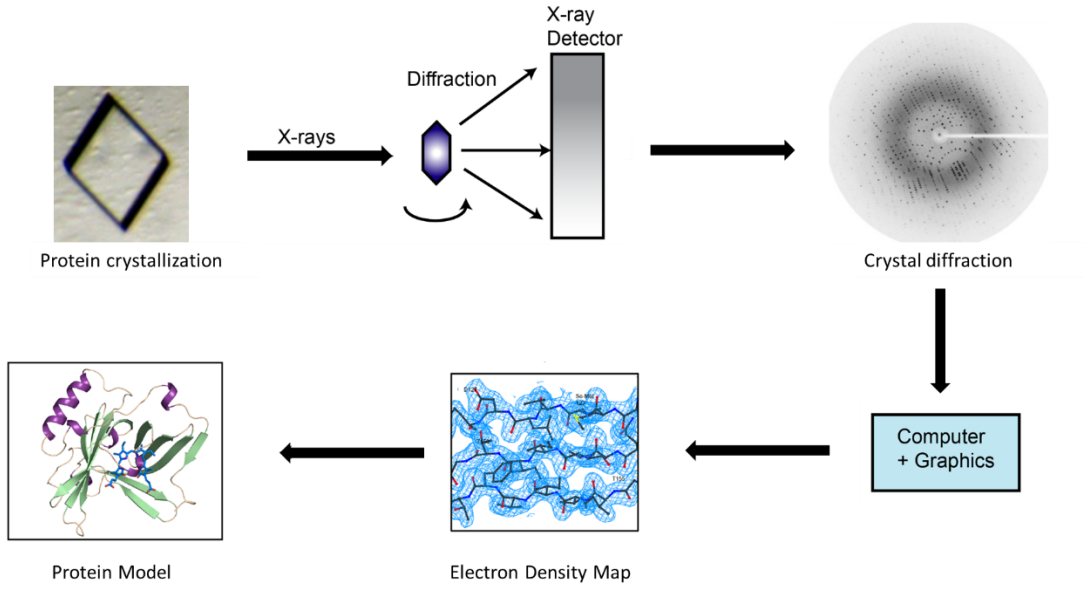


Figure 5.1- Schematic illustration of the steps to obtain the protein structure with the X-ray crystallography technique.

Crystals are repetitions of the unit cells in 3D, which are defined by 3 vectors (a, b and c) and 3 angles (α, β and γ). The molecules within the unit cell are known as asymmetric unit. A unit cell is the repeating pattern of the arrangement particle in the crystal and may have one or more asymmetric units related by symmetry operations (rotations and/without translations) [Ilari, *et al*, 2008].

There are 3 crucial requirements for the protein crystallisation: 1) the protein must be at least 90% pure to increase the chances of obtaining crystals; 2) the protein must be in a suitable buffer; 3) a precipitant solution (may be salt, organic solvent and/or polyethylene glycol) is added to the sample. The sample begins to supersaturate and small aggregates are formed, which are the nuclei of the crystals for their growth [Ilari, *et al*, 2008].

The selection of the precipitant solution is not, however, an easy task. In the supersaturation condition, the protein must be in a stable state between the solution and solid phases [Smyth, *et al*, 2000]. The time required for the protein-solution to reach the equilibrium, influences the final product: from amorphous precipitate to single crystals [Ilari, *et al*, 2008]. Apart from the precipitant solution used, the final pH, protein concentration or temperature on the crystallisation plate are also crucial for crystal growth [Smyth, *et al*, 2000].

In general, various crystallisations trials, from several available commercial kits, are tried until in one condition, there is crystal growth.

After the appearance of the crystals, a monochromatic X-ray beam is intercepted in the crystal, which in turn is scattered. Within a crystal, protein molecules are arranged in crystal planes and that in the presence of x-ray beam, scatter the x-rays. The goal here is to obtain the intensity of each reflection and the phase angle, to calculate the electron density distribution in the unit cell and thus calculate the positions of the atoms.

The reflection pattern, as described by the Braggs Law, is the constructive interference that occurs from parallel crystal planes: the reflected waves overlap, adding up together. The constructive interference holds the formula nλ=2d.SinΘ, where d is the distance from the crystal planes, Θ is the scattering angle, λ is the wavelength of the x-rays used and n is an integer (figure 5.2) [Parker, *et al*, 2003]. The radiation wavelength used must be similar to the atoms bonds distance.

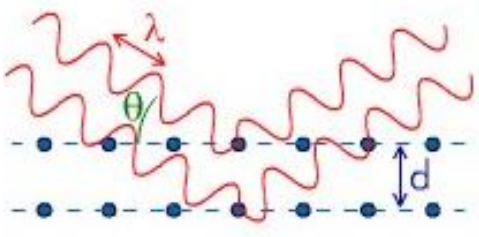


Figure 5.2- Illustration of the constructive interference, by Braggs Law. The reflected waves from parallel crystal planes overlap and are summed, holding the formula nλ=2d.SinΘ.

In the difraction experiment, the crystals are rotated to collect a maximum number of images. To each reflection spot is assigned an index of h, k, l and the diffraction intensities are calculated [Parker, *et al*, 2003]. However, the phase angle cannot be measured in the experiment, a problem known as the phase problem [Su, *et al*, 2015].

To solve the phase problem, different techniques such as multiple isomorphus replacement (MIR), multiple/single anomalous diffraction (MAD/SAD) or molecular replacement (MR) are used [Su, *et al*, 2015].

The molecular replacement method uses previously solved structures that have at least 35% of identity between the two sequences for the use of the estimated phases of the model [Abergel, *et al*, 2013]. MR involves the positioning of the model in the unit cell, which can be achieved using Patterson and maximum likelihood methods [Taylor, 2003].

In the MR method, higher the sequence identity between the model and the unknown protein structure, a greater chance of the calculated structure being correct [Abergel, *et al*, 2013]. Since MR uses the model to calculate the estimated phases, it is usually the first method chosen to solve a structure [Evans, *et al*, 2008]. However, when it fails (i.e. most of the model doesn´t fit into the electron density map), other methods mentioned above should be used.

Obtained the first electron density map calculation, this is then subjected to various refinements. The 3 positional parameters (x, y, z) and the isotropic temperature factor B for all atoms, excluding hydrogen atoms, are changed by adjusting the model, to have a closer agreement between the calculated and the observed structure factors [Ilari, *et al*, 2008].

Building the model, the validation is required for the deposition. There are several quality indicators, such as the R factor [Papageorgiou, *et al* 2014], temperature factor B [Liu, *et al*, 2014] or the Ramachandran diagram [Hollingsworth, *et al*, 2010].

The aim in this chapter was to obtain the structures of the produced CBM modules (more detailed in chapter 3) and to compare with the human malectin structure.

5.2 Materials and Methods:

5.2.1 Crystallization assay:

Theory:

There are various techniques for obtaining protein crystals. Here, we will describe the technique used, the hanging drop vapor diffusion.

In this method, a protein drop is mixed with the precipitant solution, from the reservoir, on a silanized lamella, which is then inverted over the well. This system is made air tight by using grease or silicon for the isolation (figure 5.3) [Smyth, *et al*, 2000].

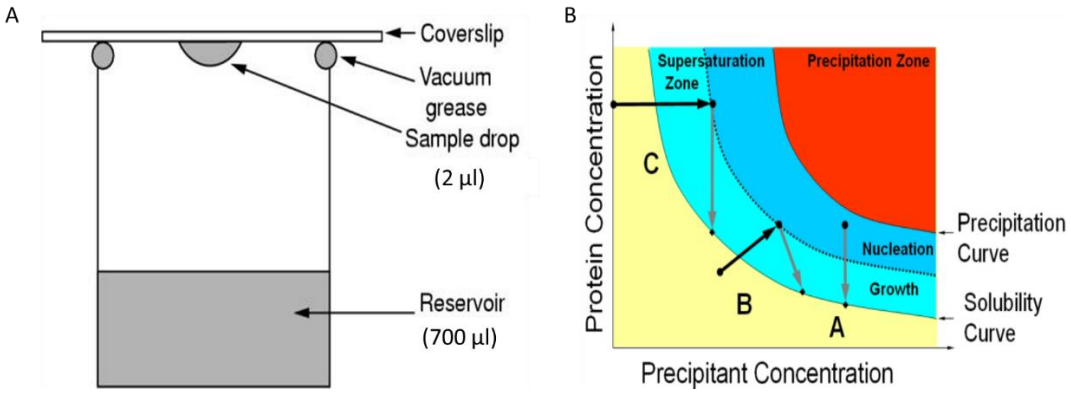


Figure 5.3- A-Schematic representation of hanging drop vapor diffusion technique; B- Phase diagram representing the concentration variation of protein and precipitant concentrations in crystallisation process.

The protein drop at the beginning has a lower concentration of the precipitant molecule than the reservoir solution [Smyth, *et al*, 2000]. The equilibrium is achieved by the passage of the water vapor from the drop to the reservoir. Slowly, the protein and precipitant concentration in the drop increases, inducing the protein nucleation. After nucleation, the nuclei begin to grow and forming crystals, until the condition reaches equilibrium.

This process for crystal growth can take hours or several weeks. As stated in the section 5.1, it is unsure which precipitant solution will cause the protein to crystalize. For new proteins, the crystallization solution has to be found and optimized to obtain high quality protein crystals, thus several screenings solutions are used, varying the salt, buffer, precipitant agent and pH have to be tested.

Experimental:

A total of 172 crystallization solutions, from the crystallization screens PEG/ION I and II, from Hampton Research, and Structure, from Molecular Dimensions (index figures 9 and 10), were manually tested on 24-well crystallisation plates for the 2 CBM modules, Bacova03100_A and Bt0996_C (production described in chapter 3) at concentrations of approximaly 25 mg/ml and at the 2 different temperatures of 4 and 20ºC.

In each reservoir, 700 µl of the respective precipitant solution was added. On the silanized glass cover slips, 1 µl of the protein solution were mixed with 1 µl of the crystallization solution. The cover slip was then inverted and placed over greased rim of the well. The 24-well crystallisation plates were then incubated at 4 and 20ºC, respectively.

5.2.2 Crystals Harvesting:

Protein crystals contains between 40-60% of solvent channels, being thus fragile and needing special care. When handled for the diffraction experiment, a harvesting solution is added (crystallization solution with higher precipitate concentration) to prevent the crystal from dissolving.

Furthermore, during the diffraction process, free radicals are formed, disrupting the integrity of the crystals. For this reason, liquid nitrogen is used to cool down the crystals. However, from this ice crystals are formed (experiment noise) and to prevent this, the crystals must be *a priori* soaked with a cryoprotectant solution [Ilari, *et al*, 2008].

First, we attempted to use the solution in the reservoir with 20% glycerol as cryoprotectant solution, however the crystals were unstable and cracked. So, the options used were: 1) the crystallization solution in the reservoir with 20% sucrose and 2) paratone.

After a crystal is put in the harvesting solution and soaked with cryo-protectant, it is stored in liquid nitrogen at -201.15 ºC until tested in the x-ray diffractometer.

5.2.3 Crystal x-ray diffraction:

The initial approach is to diffract the crystals *in house*, due to there are not certainties whether the crystals formed in the drop are protein or salt (since salt is usually included in the crystallization solutions). Typically, the diffraction of salt crystals results in a reduced number of spots with high intensity at the high-resolution shell, while the diffraction of protein crystals results in several spots with weaker intensity, in high- and low-resolution shells (figure 5.4).

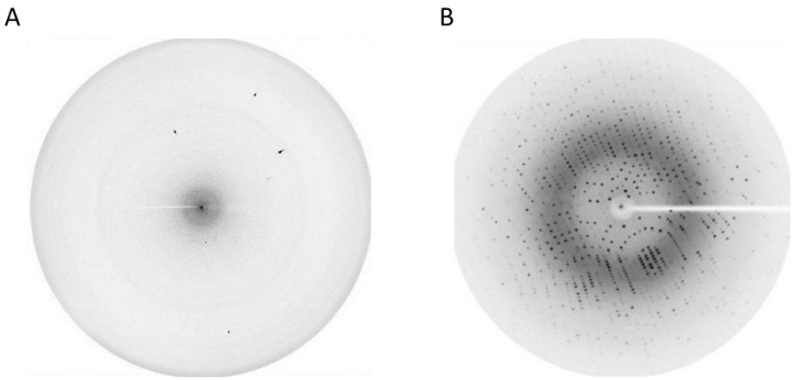


Figure 5.4- Images of x-ray diffraction pattern of a salt (A) and a protein crystal (B).

In case the crystal is of protein, x-ray diffraction *in house* has also the aim of determining if the protein crystal has a good diffraction pattern (at high resolution). Otherwise, the condition must be optimized until crystals with better quality are obtained.

Then, the next strategy is to send the crystals to synchrotron sources, in which have a more powerful x-ray beam and thus a maximum number of higher resolution diffraction images from the protein crystals can be collected [Su, *et al*, 2015].

5.3 Results and Discussion:

5.3.1 Protein crystallisation:

Of several crystallisation trials, we found one condition that enabled in 3 days the crystallisation of His-Bt0996_C module (figure 5.5).

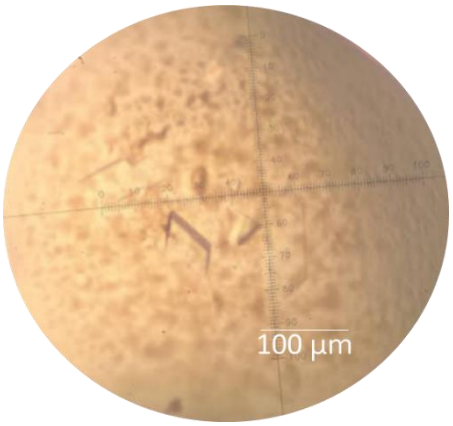


Figure 5.5- The His-BT0996_C crystals obtained at 4ºC, with 0.2 M lithium sulphate and 20% (w/v) Polyethylene glycol 3350 at pH 2.97.

Observing the drop under the microscope, it is composed of precipitated protein and of single crystals, with well-defined edges.

The condition was supposed to be composed of 0.2 M lithium sulphate, 20% (w/v) Polyethylene glycol 3350 and have a pH 6. However, the pH was measured, being 2.97. Despite several attempts, we were unable to replicate the condition using our own made solutions.

5.3.2 X-ray diffraction experiment:

The crystals were stored in liquid nitrogen (protocol described in section 5.2.2). It was observed that during the soaking with the cryoprotectant (except paratone) that the crystals began to crack, which would reveal as a major problem during data processing.

By *in house* x-ray diffraction, the diffraction pattern showed the crystals were protein (with good resolution), being thus sent to European Synchrotron Radiation Facility (ESRF).

The collected data had the best resolution of 1.5 Å (figure 5.6). However, the crystals were unstable in other cryoprotectant solutions (section 5.2.2) and began to crack, resulting that some images had several reflection spots smeared (there is a numerical reduction of reflections to be involved in electron density map calculation). On the other hand, the crystals may have not been single (as it was initially thought during microscopic observation), but a set of crystals with different orientations [Nespolo, *et al*, 2015].

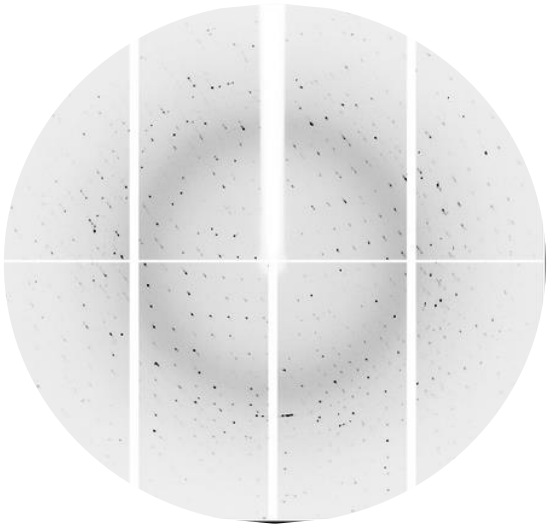


Figure 5.6- X-ray diffraction pattern from a His-Bt0996_C crystal.

The rest of the procedure to solve the structure wasn´t achieved in due time. For future work, the collected data will be processed using iMOSFLM program, which integrates all the images into a single file (mtz file). In addition, it determines the unit cell constants and refines them [Battye, *et al*, 2011]. Then, the file is proceeded to merge and scale multiple reflections into an average intensity, using the Aimless program. It has the goal of certifying the information of the space group, the data to be excluded due to the radiation damage and apply a resolution cutoff if necessary. Furthermore, it can identify: 1) possible data twinning or 2) anomalous signal [Evans, *et al*, 2013].

Scaled the data, various programs (e.g. Phaser and Phenix) will be executed using the best model for the estimation of the phases, until the generation of the best electron density map.

5.4 Conclusions:

In this chapter, we wanted to crystalize the Bacova03100 and Bt0996 proteins for structural characterization studies. Several crystallization trials were performed and Bt0996 crystals (with N-terminal His-tag) were obtained in one condition: 0.2 M Li_2SO_4 and 20% (w/v) PEG 3350 at pH of 2.97.

Despite having the appearance of a single crystal with well-defined edges under microscopic observation, the diffraction patterns showed that some reflection spots smeared/overlapped, due to the crystals cracked during freezing or because they were twinned.

A future work will be to index the collected data, estimate the phases and solve the protein structure.

Chapter 6- Global Conclusions and Future Perspectives

6. Global conclusions and future perspectives:

A total of 315 CBM modules are found and classified as belonging to the CBM57 family (to date) in the CAZy database. These modules share sequence homology with the human malectin and are present in all domains of life. In this thesis, we began first to perform bioinformatic studies for a better understanding of the malectin evolution.

The first step was to predict the protein architecture of the associated modules, using InterProScan. It has been observed that CBM57 modules in the Animalia Domain are individualized (like the human malectin), while in plants they are associated with kinases and in bacteria they are usually appended to a TolB-like transporter or to a catalytic module such as GH2, S8-S53 peptidases, pectin lyases, PapD-like or galactose oxidases. The second step was to form groups based on the Domain of life to which they belong. In the Bacteria domain, the modules were grouped accordingly to the type of modules they were appended. Then to visualize the conservation level of structural elements as well as the putative binding-sites, their amino acids sequences were aligned using ClustalOmega. As in Schallus et al., 2008, our alignments predict that CBM57 modules from eukaryotes (excluding plants) to have the same specificity as the human malectin. The exception is the module from *C.parvum*, which has 2 putative binding-residues replaced. For the remaining members of the CBM57 family, we also expect them to have different specificities from the human malectin. CBM57 modules appended to TolB-like and/or PKA have among them a high level of conservation, thus we speculate that a convergent evolution occurred for these modules contrary to what seems to have occurred with other bacterial modules. This was particularly noticed for the CBM57 module appended to GH2, which showed a higher level of divergence, being thus expected to have various specificities.

To verify this assumption, a total of 7 CBM modules (members of CBM57 family, but also other homologous, such as CBM6 and CBM35) found in 2 bacteria species belonging to the human gut - *Bacteroides ovatus* and *Bacteroides thetaiotaomicron* were attempted for expression and biochemical characterization. Furthermore, this work aimed at a better understanding of the mechanisms used in these 2 species for the degradation of complex polysaccharides.

For the expression tests, we used 2 strains and more than 72 different recombinant protein expression conditions were tested. Furthermore, we have re-cloned the recombinant DNAs from the pNZY vector to pET28 vector, changing the His-tag position in order to increase the probability of expressing these modules. From the initial 7 recombinant DNAs, 5 were successfully re-cloned. Despite this effort, just 2 CBM modules expressed in soluble form: a CBM35 module from *B.ovatus* (Bacova_A) and CBM module from *B.thetaiotaomcron* (Bt0996_C). The later could not be assigned to a specific family, but shared sequence homology with families 6, 35 and 57.

The CBM module from *B.thetaiotaomicron* expressed in the 2 constructs, thus we produced both, since it could influence the biochemical characterization. At this stage, differences were observed with the C-terminal His-tag, which improved the expression of the module, but the purification process required an additional chromatography.

To identify the specificity of these 2 modules, 2 glycan microarrays were performed: 1) a manual microarray to initially identify the type of polysaccharide to which they bind and 2) a robotic microarray to determine more specifically the epitope recognized.

It was known that the *B.thetaiotaomicron* can be grown in a medium with just RG II as energy source, thus the manual microarray was composed mainly of pectins, with few fungal polysaccharides and fractions from microalga *N.oculata*. It revealed that Bt0996_C recognizes some pectins such as galacturonate and PGA from citrus. It was also shown that this module doesn´t bind to highly methylated pectins, since there was a lack of interaction with high-methylated galacturonate. The robotic microarray, with a higher number of glycan probes to be tested, showed that this module can also interact with the fraction of *P.biglobosa* and a fraction (50WSnFI-Se-EI) of *S.nigra*. Furthermore, in this microarray we tested the module with the His-tag in 2 positions and slight differences were observed. The His-Bt0996_C recognized a fraction (Et50) of *N.oculata* both in the manual and robotic microarrays, but for Bt0996_C-His this interaction wasn´t present. This showed that the His-tag position had influence in our results, therefore the interaction of His-Bt0996_C with the *N.oculata* fraction was non-specific. Although

we have no certainty about the epitope recognition, we speculate that it may be rhamnose and galacturonic acid linked by α-(1-3) linkage, a saccharide sequence observed in RG II that isn´t present in the RG I. The Bt0996_C appears to recognize a different saccharide sequence from the joined catalytic modules

For the Bacova03100_A, on the other hand, no information was known about the type of polysaccharide to which it could bind to. We attempted to include this module in both glycan microarrays, but no interactions were observed. This may be due to the fact that the module interacts with other type of polysaccharides.

Since the glycans samples used were from natural sources, they are often heterogeneous despite the performed purification processes. Characterized proteins are simultaneously tested with our proteins in order to validate the glycan microarray analysis. Furthermore, it is also important to test the antibodies involved in the detection system. As shown in the robotic microarray, there were non-specific interactions for carrageenan kappa, lota and lamda, thus excluded in our analysis.

For structural characterization studies, crystallization trials were performed and crystals of Bt0996_C were obtained in the following condition: 0.2 M Lithium sulphate and 20% (w/v) PEG 3350 at pH of 2.97. However, we were unable to replicate the crystallization solution. This led to a problem with the chosen cryoprotectant and the crystals became unstable and cracked during the freezing process. The diffraction *in house* has proven to be crucial in knowing if the crystals obtained were protein, before sending them to a more powerful x-rays beam source, the synchrotron. The Bt0996_C crystals had a best resolution of 1.5 Å, but several reflection spots were smeared and overlapped, showing the possibility of being initially twinned.

For future work, further attempts will be made in order to solve the structure of Bt0996-C. Moreover, new glycan microarrays will be performed with different new fragments of RG II in order to discover the epitope recognition of Bt0996_C. Hence, co-crystallization with the oligosaccharide will be performed for a better understanding of the binding-site and to compare twith the structure of human malectin with its ligand.

References:

Abergel, Chantal. "Molecular replacement: tricks and treats." *Acta Crystallographica Section D: Biological Crystallography* 69.11 (2013): 2167-2173.

Barsett, Hilde, et al. "Comparison of carbohydrate structures and immunomodulating properties of extracts from berries and flowers of Sambucus nigra L." *European Journal of Medicinal Plants* 2.3 (2012): 216.

Battye, T. Geoff G., et al. "iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM." *Acta Crystallographica Section D: Biological Crystallography* 67.4 (2011): 271-281.

Bledzki, A. K., and Jochen Gassan. "Composites reinforced with cellulose based fibres." *Progress in polymer science* 24.2 (1999): 221-274.

Blixt O, Razi N. Chemoenzymatic synthesis of glycan libraries. Methods Enzymol. 2006;415:137–53.

Blixt O, Head S, Mondala T, Scanlan C, Huflejt M, et al. Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. Proc Natl Acad Sci U S A. 2004;101:17033–8.

Boraston, Alisdair B., et al. "Carbohydrate-binding modules: fine-tuning polysaccharide recognition." *Biochemical Journal* 382.3 (2004): 769-781.

Bornhorst, Joshua A., and Joseph J. Falke. "[16] Purification of proteins using polyhistidine affinity tags." *Methods in enzymology* 326 (2000): 245-254.

Burton, Rachel A., and Geoffrey B. Fincher. "(1, 3; 1, 4)-β-d-Glucans in cell walls of the Poaceae, lower plants, and fungi: a tale of two linkages." *Molecular Plant* 2.5 (2009): 873-882.

Carroll G, Wang D, Turro N, Koberstein J. Photochemical micropatterning of carbohydrates on a surface. Langmuir. 2006;22:2899–905.

Demain, Arnold L., and Preeti Vaishnav. "Production of recombinant proteins by microbes and higher organisms." *Biotechnology advances* 27.3 (2009): 297-306.

Deprez, Paola, Matthias Gautschi, and Ari Helenius. "More than one glycan is needed for ER glucosidase II to allow entry of glycoproteins into the calnexin/calreticulin cycle." *Molecular cell* 19.2 (2005): 183-195.

Evans, Philip, and Airlie McCoy. "An introduction to molecular replacement." *Acta Crystallographica Section D: Biological Crystallography* 64.1 (2008): 1-10.

Evans, Philip R., and Garib N. Murshudov. "How good are my data and what is the resolution?." *Acta Crystallographica Section D: Biological Crystallography* 69.7 (2013): 1204-1214.

Fernández-Castané, Alfred, et al. "Evidencing the role of lactose permease in IPTG uptake by Escherichia coli in fed-batch high cell density cultures." *Journal of biotechnology* 157.3 (2012): 391-398.

Galli, Carmela, et al. "Malectin participates in a backup glycoprotein quality control pathway in the mammalian ER." *PLoS One* 6.1 (2011): e16304.

Garibyan, Lilit, and Nidhi Avashia. "Research techniques made simple: polymerase chain reaction (PCR)." *The Journal of investigative dermatology* 133.3 (2013): e6.

Gorshkova, T. A., et al. "Formation of plant cell wall supramolecular structure." *Biochemistry (Moscow)* 75.2 (2010): 159-172.

Heimburg-Molinaro, Jamie, et al. "Preparation and analysis of glycan microarrays." *Current protocols in protein science* (2011): 12-10.

Henshaw, Joanna L., et al. "The family 6 carbohydrate binding module CmCBM6-2 contains two ligand-binding sites with distinct specificities." *Journal of Biological Chemistry* 279.20 (2004): 21552-21559.

Hollingsworth, Scott A., and P. Andrew Karplus. "A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins." *Biomolecular concepts* 1.3-4 (2010): 271-283.

Hong, Paula, Stephan Koza, and Edouard SP Bouvier. "A review size-exclusion chromatography for the analysis of protein biotherapeutics and their aggregates." *Journal of liquid chromatography & related technologies* 35.20 (2012): 2923-2950.

Hudson, Kieran L., et al. "Carbohydrate–aromatic interactions in proteins." *Journal of the American Chemical Society* 137.48 (2015): 15152-15160.

Ilari, Andrea, and Carmelinda Savino. "Protein structure determination by x-ray crystallography." *Bioinformatics: Data, Sequence Analysis and Evolution* (2008): 63-87.

Jones, Louise, Graham B. Seymour, and J. Paul Knox. "Localization of pectic galactan in tomato cell walls using a monoclonal antibody specific to (1 [->] 4)-[beta]-D-galactan." *Plant Physiology* 113.4 (1997): 1405-1412.

Jones, Philip, et al. "InterProScan 5: genome-scale protein function classification." *Bioinformatics* 30.9 (2014): 1236-1240.

Kuhn, Heiko, and Maxim D. Frank-Kamenetskii. "Template-independent ligation of single-stranded DNA by T4 DNA ligase." *The FEBS journal* 272.23 (2005): 5991-6000.

Kumar, Sudhir, Glen Stecher, and Koichiro Tamura. "MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets." *Molecular biology and evolution* 33.7 (2016): 1870-1874.

Lehle, Ludwig, Sabine Strahl, and Widmar Tanner. "Protein glycosylation, conserved from yeast to man: a model organism helps elucidate congenital human diseases." *Angewandte Chemie International Edition* 45.41 (2006): 6802-6818.

Liu Y, Feizi T, Campanero-Rhodes M, Childs R, Zhang Y, et al. Neoglycolipid probes prepared via oxime ligation for microarray analysis of oligosaccharide-protein interactions. Chem Biol. 2007;14:847–59.

Liu, Qian, Zhenhua Li, and Jinyan Li. "Use B-factor related features for accurate classification between protein binding interfaces and crystal packing contacts." *BMC bioinformatics* 15.16 (2014): S3.

Lodish, Harvey, et al. "DNA cloning with plasmid vectors." (2000). Molecular Cell biology, 4th edition, Section 7.1

Lynch, Susan V., and Oluf Pedersen. "The human intestinal microbiome in health and disease." *New England Journal of Medicine* 375.24 (2016): 2369-2379.

Marcus, Susan E., et al. "Restricted access of proteins to mannan polysaccharides in intact plant cell walls." *The Plant Journal* 64.2 (2010): 191-203.

Martens, Eric C., et al. "Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts." *PLoS biology* 9.12 (2011): e1001221.

Mori, Sadao, and Howard G. Barth. *Size exclusion chromatography*. Springer Science & Business Media, 2013.

Nakamura, Aline M., Alessandro S. Nascimento, and Igor Polikarpov. "Structural diversity of carbohydrate esterases." *Biotechnology Research and Innovation* (2017).

Naumoff, D. G. "Hierarchical classification of glycoside hydrolases." *Biochemistry (Moscow)* 76.6 (2011): 622-635.

Nei, Masatoshi, and Jianzhi Zhang. "Evolutionary distance: estimation." *eLS* (2006).

Nespolo, Massimo. "Tips and traps on crystal twinning: how to fully describe your twin." *Crystal Research and Technology* 50.5 (2015): 362-371.

Ndeh, Didier, et al. "Complex pectin metabolism by gut bacteria reveals novel catalytic functions." *Nature* 544.7648 (2017): 65-70.

Palma, A.S. *et al.*, 2015. Unravelling glucan recognition systems by glucome microarrays using the designer approach and mass spectrometry. *Molecular and Cellular Proteomics*, (5), pp.3105–3117.

Papageorgiou, Anastassios C., and Jesse Mattsson. "Protein structure validation and analysis with X-ray crystallography." *Protein Downstream Processing: Design, Development and Application of High and Low-Resolution Methods* (2014): 397-421.

Parker, M. W. "Protein structure from X-ray diffraction." *Journal of biological physics* 29.4 (2003): 341-362.

Pavlopoulos, Georgios A., et al. "A reference guide for tree analysis and visualization." *BioData mining* 3.1 (2010): 1.

Perez, S. "The symbolic representation of monosaccharides in the age of glycobiology." (2014): 1-19.

Rillahan, Cory D., and James C. Paulson. "Glycan microarrays for decoding the glycome." *Annual review of biochemistry* 80 (2011): 797-823.

Rosano, Germán L., and Eduardo A. Ceccarelli. "Recombinant protein expression in Escherichia coli: advances and challenges." *Frontiers in microbiology* 5 (2014).

Roy, S., and V. Kumar. "Practical approach on SDS PAGE for separation of protein." *Int. J. Sci. Res* 3.8 (2014): 955-960.

Schallus, Thomas, et al. "Malectin: a novel carbohydrate-binding protein of the endoplasmic reticulum and a candidate player in the early steps of protein N-glycosylation." *Molecular biology of the cell* 19.8 (2008): 3404-3414.

Schallus, Thomas, et al. "Analysis of the specific interactions between the lectin domain of malectin and diglucosides." *Glycobiology* 20.8 (2010): 1010-1020.

Scheller, H.V. & Ulvskov, P., 2010. Hemicelluloses. *Annual Review of Plant Biology*, 61(1), pp.263–289.

Shoseyov, Oded, Ziv Shani, and Ilan Levy. "Carbohydrate binding modules: biochemical properties and novel applications." *Microbiology and molecular biology reviews* 70.2 (2006): 283-295.

Shipman, Joseph A., James E. Berleman, and Abigail A. Salyers. "Characterization of Four Outer Membrane Proteins Involved in Binding Starch to the Cell Surface of Bacteroides thetaiotaomicron." *Journal of bacteriology* 182.19 (2000): 5365-5372.

Shreiner, Andrew B., John Y. Kao, and Vincent B. Young. "The gut microbiome in health and in disease." *Current opinion in gastroenterology* 31.1 (2015): 69.

Sievers, Fabian, et al. "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega." *Molecular systems biology* 7.1 (2011): 539.Robert, X. and Robert, X. and Gouet, P. (2014) "Deciphering key features in protein structures with the new ENDscript server". *Nucl. Acids Res*. **42**(W1), W320-W324.

Smoot J, Demchenko A. Oligosaccharide synthesis: from conventional methods to modern expeditious strategies. Adv Carbohydr Chem Biochem. 2009;62:161–250

Soskine, Misha, and Dan S. Tawfik. "Mutational effects and the evolution of new protein functions." *Nature Reviews Genetics* 11.8 (2010): 572-582.

Stevens, Raymond C. "Design of high-throughput methods of protein production for structural biology." *Structure* 8.9 (2000): R177-R185.

Su, Xiao-Dong, et al. "Protein crystallography from the perspective of technology developments." *Crystallography reviews* 21.1-2 (2015): 122-153.

Taylor, Garry. "The phase problem." *Acta Crystallographica Section D: Biological Crystallography* 59.11 (2003): 1881-1890.

Taylor Maureen, E., and K. Drickamer. *Introduction to Glycobiology*. Oxford University Press, 2003.

Thomas, François, et al. "Environmental and gut bacteroidetes: the food connection." *Frontiers in microbiology* 2 (2011).

Van Pelt-Verkuil, Elizabeth, Alex van Belkum, and John P. Hays. "Analysis of PCR Amplification Products." *Principles and Technical Aspects of PCR Amplification* (2008): 141-182.

Varki, Ajit, and John B. Lowe. "Biological roles of glycans." (2009) chapters.

Varki A, Cummings R, Esko J, Freeze H, Stanley P, et al. Essentials of Glycobiology. New York: Cold Spring Harbor; 2009. p. 830

Wang, L. *et al.*, 2014. Cross-platform comparison of glycan microarray formats. *Glycobiology*, 24(6), pp.507–517.

Wang, Yufeng, et al. "Specificities of Ricinus communis agglutinin 120 interaction with sulfated galactose." *FEBS letters* 585.24 (2011): 3927-3934.
Wang D, Liu S, Trummer B, Deng C, Wang A. Carbohydrate microarrays for the recognition of cross-reactive molecular markers of microbes and host cells. Nat Biotechnol. 2002;20:275–81.

Willats W, Rasmussen S, Kristensen T, Mikkelsen J, Knox J. Sugar-coated microarrays: a novel slide surface for the high-throughput analysis of glycans. Proteomics. 2002;2:1666–71.

Woodman, Michael E. "Direct PCR of intact bacteria (colony PCR)." *Current protocols in microbiology* (2008): A-3D.

Xiong, Jin. *Essential bioinformatics*.Chapters 5, 10 and 11. Cambridge University Press, 2006.

Yang, Ziheng, and Bruce Rannala. "Molecular phylogenetics: principles and practice." *Nature Reviews Genetics* 13.5 (2012): 303-314.

www.cazy.org, last access 03.09.2017

Index

Index table 1- Carbohydrate binding modules (CBMs) for production to biochemical characterization, by glycan microarray and X-ray Chrystallography. CBMs produced and under study are highlighted in bold, respective recombinant protein sequence, protein identification, family, molecular weight, base pairs, extinction coefficient and isoelectric point are described.

Microorganism	Molecular Architecture	Recombinant DNA sequence	Protein ID	Family	Molecular weight (Kda)	Base pairs	Extinction coefficient (M^{-1})	Theoretical Isoelectric point
B.ovatus	SP-**GBD(CBM)**-CBM(GH2)-GH2-UNK-CBM57	MGSSHHHHHHSSGPQQGLRCGQRVTQTINDG WKFSLFEGDASTADFDVSGWTDVSIPHTWNAK DAEDEIPGYFRGKGWYRKAVTVEELIVGQRVYL CFEGANQETNVFVNGKLVGNHKGGYSAFTFDV TDYVHTGRNLVAVSVDNSYNPDIAPLSADFTFFG GLYRDVYLVYTS	Bacova03100_A	35	19.38	465	35535	5.62
B. thetaiotaomicron	SP-GH?-GBD/CBM-GH2-UNK-MAL-**CBM(6/35/57)**	MGSSHHHHHHSSGPQQGLRYEAETATLKGKFR KKEHRKQTGVFFDKGKGNSIEWNISTGLAQVYA LRFKYMNTTGKPMPVLMKFIDSKGVVLKEDILTF PETPDKWKMMSTTTGTFINAGHYKVLLSAENM DGLAFDALDI	Bt0996_C	6/35/57	15.89	366	16960	9.43

Index table 2- List of all saccharide probes included in the robot glycan microarray. The microarray is comprised of polysaccharide samples from different sources, representative of major sequences found in plant cell walls, and a few selected sequence-defined oligosaccharides as acidic or mammals glycans. The probes are grouped according to the printing layout (set position).

Pos[a]	Probe	Predominant oligosaccharide sequence/ monosaccharide composition
1	Pectin-1 (*Terminalia macroptera*)	Under characterization
2	Pectin-2 (*Terminalia macroptera*)	Under characterization
3	Pectin-3 (*Terminalia macroptera*)	Under characterization
4	Pectin-4 (*Terminalia macroptera*)	Under characterization
5	Pectin-5 (*Terminalia macroptera*)	Under characterization
6	Pectin-6 (*Terminalia macroptera*)	Under characterization
7	Pectin-7 (*Terminalia macroptera*)	Under characterization
8	Pectin-8 (*Terminalia macroptera*)	Under characterization
9	Pectin-9 (*Terminalia macroptera*)	Under characterization
10	Pectin-10 (*Terminalia macroptera*)	Under characterization
11	Pectin-11 (*Terminalia macroptera*)	Under characterization
12	Pectin-12 (*Terminalia macroptera*)	Under characterization
13	Pectin-13 (*Terminalia macroptera*)	Under characterization
14	Carragean k (*red algae*)	Mixed linked α-1,3/β- 1,4 galactoses
15	Carragean I (*red algae*)	Mixed linked α-1,3/β- 1,4 galactoses
16	Carragean λ (*red algae*)	Mixed linked α-1,3/β- 1,4 galactoses
17	Polygalacturonic Acid (PGA) Citrus	α 1-4 galacturonic acid backbone
18	Galacturonate Low Methylated Apple	α 1-4 galacturonic acid backbone
19	Galacturonate Low Methylated Citrus	α 1-4 galacturonic acid backbone
20	Pectic Galactan Lupin	β1-4 Gal (Gal: Ara: Rha: Xyl: GalUA = 77: 14: 3: 0.6: 5.4)

21	Pectic Galactan Potato	β1-4 Gal (Gal: Ara: Rha: GalUA = 78: 9: 4: 9)
22	Galactan Lupin	β1-4 Gal (Gal:Ara:Rha:Xyl:other sugars = 82 : 5.8 : 5.1 : 1.4 : 5.7, Galacturonic acid 14.6%.
23	Rhamnogalacturonan Soy bean	Mixed-linked α-(1-2;1-4) rhamnose and galacturonic acid backbone
24	Fraction Et50 (SEC:630 kDa) *N. oculata*	Under characterization
25	50WSnFl-S2 (*Sambucus nigra*) (ID17)	Ara (28%), Rha (4.4%), Xyl (1.3%), Man (1%), Gal (19.2%), Glc (2%), GlcA (0.4%), GalA (42.3%), 4-O-Me-GlcA (1.4%)
26	100WSnFl-S2 (*Sambucus nigra*) (ID18)	Ara (18.2%), Rha (16.8%), Xyl (3.4%), Man (0.5%), Gal (17.8%), Glc (2.8%), GlcA (0.3%), GalA (40.5%), 4-O-Me-GlcA (0.9%)
27	50WSnFl-S2-EI (*Sambucus nigra*) (ID19)	Ara (29.5%), Rha (14.3%), Fuc (0.4%), Xyl (1.8%), Man (2.0%), Gal (25.5%), Glc (3.6%), GlcA (2.3%), GalA (17.9%), 4-O-Me-GlcA (2.7%)
28	Al50W-I-2 (*Allium cepa* mild) (ID20)	Under characterization
29	IOI-WAc *(Inonotus obliquus)* (ID4)	Under characterization
30	IOI-WN *(Inonotus obliquus)* (ID5)	Under characterization
31	CC1P1(*Cola cordifolia bark)* (ID11)	Ara (trace), Rha (32%), Gal (31%), Glc (2%), GalA (35%)
32	SnFl50-S2 (*Sambucus nigra*) (ID16)	Ara (19.4%), Rha (5.3%), Xyl (0.7%), Man (1.1%), Gal (22.9%), Glc (2.8%), GlcA (2.1%), GalA (44.7%), 4-O-Me-GlcA (1%)
33	BP-II *(Biophytum petersianum*) (ID 1)	Ara (5.1%), Rha (8.2%), Fuc (0.5%), 2-Me-Fuc (trace), Xyl (6.3%), 2-Me-Xyl (trace), Man (0.7%), Gal (8.3%), Glc (4.4%), GlcA (1.3%), GalA (65.1%)
34	GOA1 (*Glinus oppositifolius*) (ID2)	Ara (26.4%), Rha (4.2%), Xyl (3.9%), Man (4.3%), Gal (42.9%), Glc (3.5%), GalA (12.1%), 4-O-Me-GlcA (2.9%)
35	GOA2 (*Glinus oppositifolius*) (ID3)	Ara (5.5%), Rha (10.3%), Fuc (1.3%), Xyl (0.5%), Man (0.6%), Gal (9.7%), Glc (3.3%), GalA (68.3%), 4-O-Me-GlcA (0.4%)
36	Vk100-Fr.I (*Vernonia kotschyana*) (ID6)	Ara (2%), Rha (1%), Fru (83%), Gal (2%), Glc (3%), GalA (1%)
37	Ctw-A1 (*Cochlospermum tinctorium*) (ID7)	Ara (16.3%), Rha (17.9%), Man (1.8%), Gal (45.8%), Glc (4%), GlcA (8.8%), GalA (5.8%), Fru (4.9%)
38	Oc50A1.IA (*Opilia celtidifolia*) (ID8)	Ara (38.9%), Rha (4.2%), Man (5.8%), Gal (30.9%), Glc (5.4%), GlcA (trace), GalA (11.5%), 4-O-Me-GlcA (3.3%)
39	LPS3 (*Tilia cordata)* (ID9)	Under characterization
40	CC1 (*Cola cordifolia bark)* (ID10)	Ara (3.7%), Rha (22.1%), Gal (20.2%), Glc (0.5%), GalA (29.6%), 2-O-Me-Gal (6.5%), 4-O-Me-GlcA (17.4%)
41	CC2 (*Cola cordifolia bark)* (ID12)	Ara (37.2%), Rha (8.5%), Gal (31.3%), Glc (1.1%), GalA (11.5%), GlcA (3.4%), 2-O-Me-Gal (0.4%), 4-O-Me-GlcA (6.6%)
42	CC3 (*Cola cordifolia bark)* (ID13)	Ara (3.0%), Rha (22.8%), Gal (17.3%), Glc (1%), GalA (32.8%), 4-O-Me-GlcA (17.8%), 2-O-Me-Gal (5.3%)
43	PBS100-II (*Parkia biglobosa)* (ID15)	Ara (21.2%), Rha (7.3%), Xyl (0.2%), Gal (18%), Glc (6.1%), GalA (30.1%), GlcA (10.5%), 4-O-Me-GlcA (1.3%)
44	Fraction Et85 (SEC:65 kDa) *N. oculata*	Under characterization
45	Fraction Et85-1 (Ion exchange: 19 kDa) *N. oculata*	Under characterization
46	Fraction Et85-2 (Ion exchange: 47 and 448 kDa) *N. oculata*	Under characterization

47	Et85-3 (Ion exchange: 14 and 339 kDa) *N. oculata*	Under characterization
48	Fraction EtSU (SEC: 26 kDa) *N. oculata*	Under characterization
49	Xylan (Plum) (X3)	Rha (8%), Fuc (0%), Ara (13%), Xyl (69%), Man (2%), Gal (3%), Glc (1%), Ur Ac (3%)
50	Xylan (Plum) (X4)	Rha (0%), Fuc (0%), Ara (11%), Xyl (59%), Man (0%), Gal (2%), Glc (4%), Ur Ac (25%)
51	Xylan (Plum) (X5)	Rha (2%), Fuc (2%), Ara (10%), Xyl (63%), Man (0%), Gal (5%), Glc (5%), Ur Ac (13%)
52	Xyloglucan (Plum) (XG3)	Rha (2%), Fuc (6%), Ara (5%), Xyl (44%), Man (4%), Gal (12%), Glc (25%), Ur Ac (2%)
53	Xyloglucan (Plum) (XG4)	Rha (2%), Fuc (5%), Ara (3%), Xyl (36%), Man (7%), Gal (11%), Glc (29%), Ur Ac (8%)
54	Xyloglucan (Plum) (XG5)	Rha (4%), Fuc (5%), Ara (5%), Xyl (35%), Man (5%), Gal (11%), Glc (28%), Ur Ac (4%)
55	CSA (Sigma-Aldrich)	mixed-linked-β-(1-3,1-4) glucuronic acid and N-acetyl-galactosamine-4-sulfate
56	CSB (Sigma-Aldrich)	β1-3 L-iduronic acid and N-acetyl-galactosamine-4-sulfate
57	CSC (Sigma-Aldrich)	mixed-linked-β-(1-3,1-4) glucuronic acid and N-acetyl-galactosamine-6-sulfate
58	HA (Sigma-Aldrich)	mixed-linked-β-(1-3,1-4) glucuronic acid and N-acetyl-galactosamine
59	Galactoglycolipid from *N. oculata*	Under characterization
60	LCC (*Cola cordifolia leaf)* (ID14)	Under characterization

[a] Position matching the original microarray set.

Index table 3- List of all characterized proteins included for the quality control validation of the glycan microarrays, with their respective preference oligosaccharide sequence recognition.

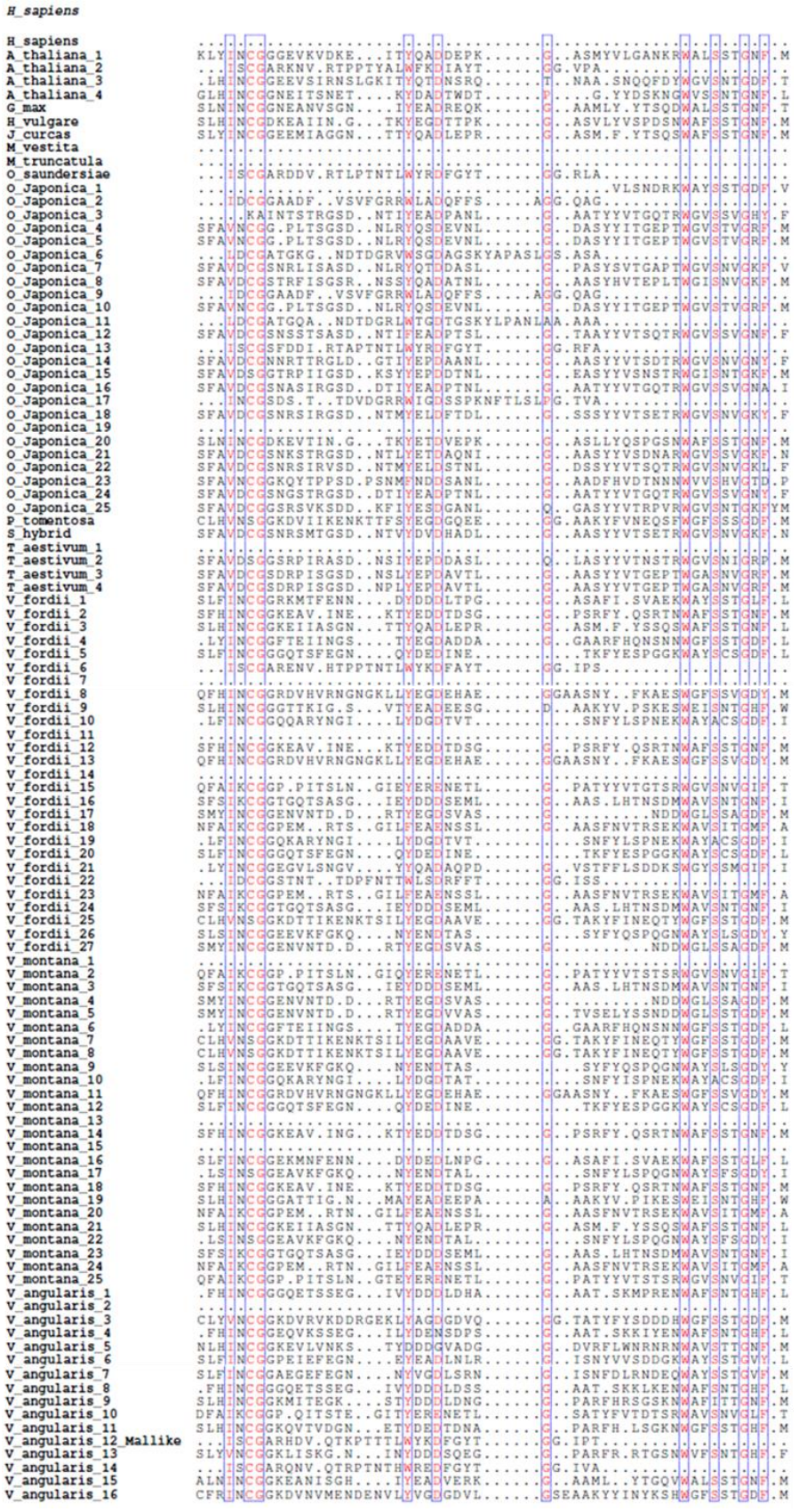
Protein Name	Preference Oligosaccharide sequence recognition	Reference
Anti-(1,3/1,4)-β-D-Glucan	Mixed linked β 1,3/1,4 glucoses	[Burton, *et al*, 2009]
Anti-1,4-β-D-Galactan	β 1-4 galactoses linkages	[Jones, *et al*, 1997]
LM21	β1-4 linked mannose and glucose	[Marcus, *et al*, 2010]
Concanavalin A	Terminal β-1-4 linked mannoses	[Wang, *et al*, 2014]
Ricinus communis agglutinin I	Terminal β-linked galactoses	[Wang, *et al*, 2011]
CmCBM6-2	Mixed linked β 1,3/1,4 glucoses	[Henshaw, *et al*, 2004]

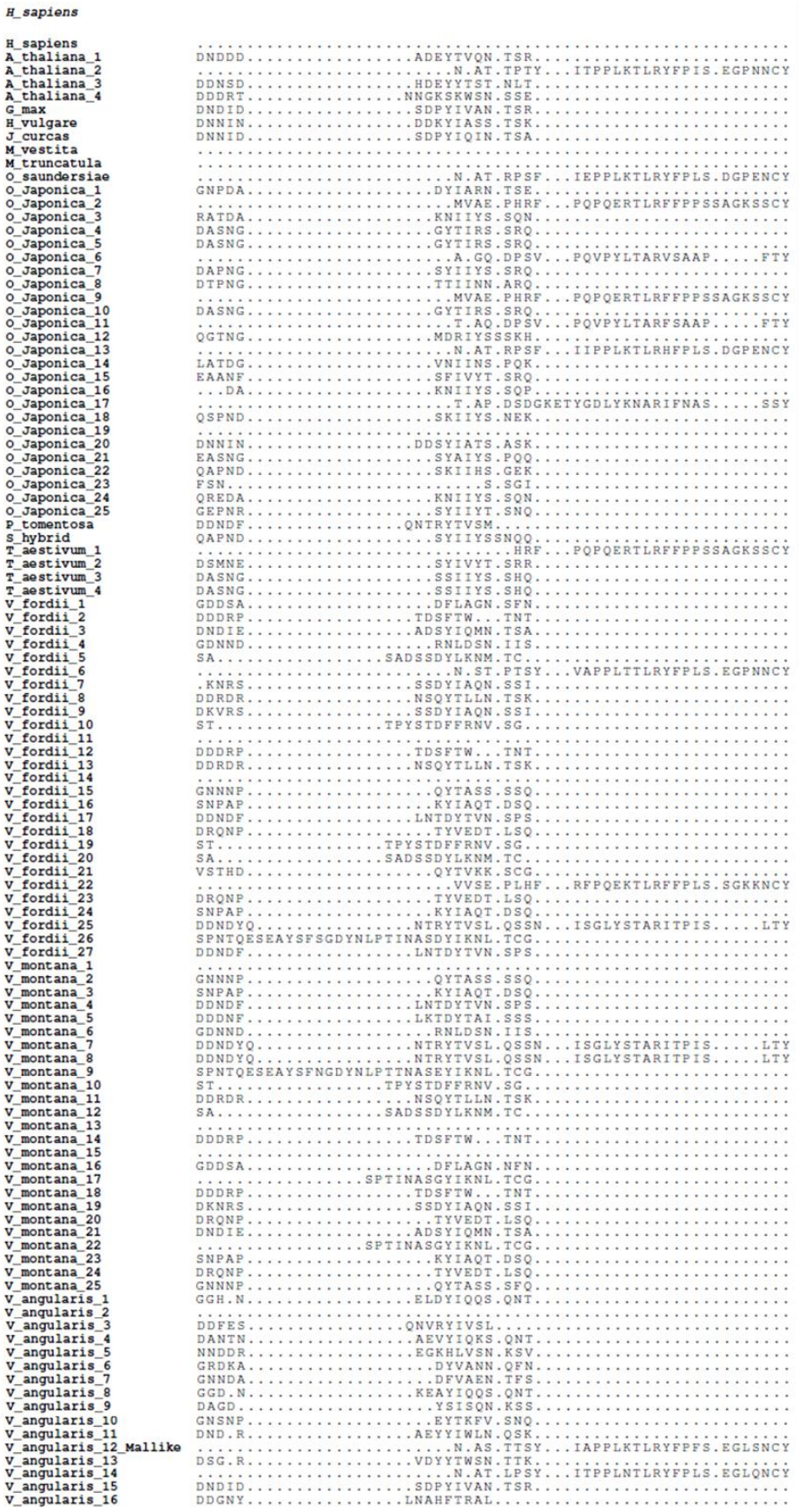
Index table 4- Information of the solutions prepared for glycan microarrays.

Protein: Concentration: Antibody: Biodetector: Blocker/ Diluent	manti-His 15 μg/ml 15 μg/ml N/A 0.02% Casein, 1% BSA in HBS, 5 mM $CaCl_2$
Protein: Concentration: Antibody: Biodetector: Blocker/ Diluent Location	BI anti mouse IgG 15 μg/ml N/A N/A 0.02% Casein, 1% BSA in HBS, 5 mM $CaCl_2$
Protein: Concentration: Antibody: Biodetector: Blocker/ Diluent	His-BT0996_C 25 μg/ml manti-His 75 μg/ml BI anti mouse IgG 75μg/ml 0.02% Casein, 1% BSA in HBS, 5 mM $CaCl_2$ 4ºC Lab.415
Protein: Concentration: Antibody: Biodetector: Blocker/ Diluent	BT0996_C-His 25 μg/ml manti-His 75 μg/ml BI anti mouse IgG 75μg/ml 0.02% Casein, 1% BSA in HBS, 5 mM $CaCl_2$
Protein: Concentration: Antibody: Biodetector: Blocker/ Diluent	BACOVA03100_A 25 μg/ml manti-His 75 μg/ml BI anti mouse IgG 75μg/ml 0.02% Casein, 1% BSA in HBS, 5 mM $CaCl_2$
Protein: Concentration: Antibody: Biodetector: Blocker/ Diluent	*Ricinus Communis* Agglutinin I (RCA_{120}) 2 μg/ml N/A N/A 1% Casein 1:50 1% BSA, 5 mM $CaCl_2$
Protein: Concentration: Antibody: Biodetector: Blocker/ Diluent	Anti (1-4)-β-D-Galactan (Rat IgG) 1/10 Dilution BI anti rat IgG 3μg/ml (800 μg/ml) N/A 1% Casein 1:50 1% BSA, 5 mM $CaCl_2$
Protein: Concentration: Antibody: Biodetector: Blocker/ Diluent	Anti-(1-3;1-4)-beta-D-glucan (Mouse IgG) 10 μg/ml BI anti-mouse IgG 3 μg/ml N/A 1% Casein 1:50 1% BSA, 5 mM $CaCl_2$

Protein: Concentration: Antibody: Biodetector: Blocker/ Diluent	BI-Concanavalin A (Con A) 2 μg/ml N/A N/A 3% BSA, 5 mM CaCl2
Protein: Concentration: Antibody: Biodetector: Blocker/ Diluent	CmCBM6-2 10 μg/ml Mouse anti-His 30 μg/mL BI anti-mouse IgG 30 μg/mL 1% BSA, 0.02% Casein in HBS, 5mM $CaCl_2$

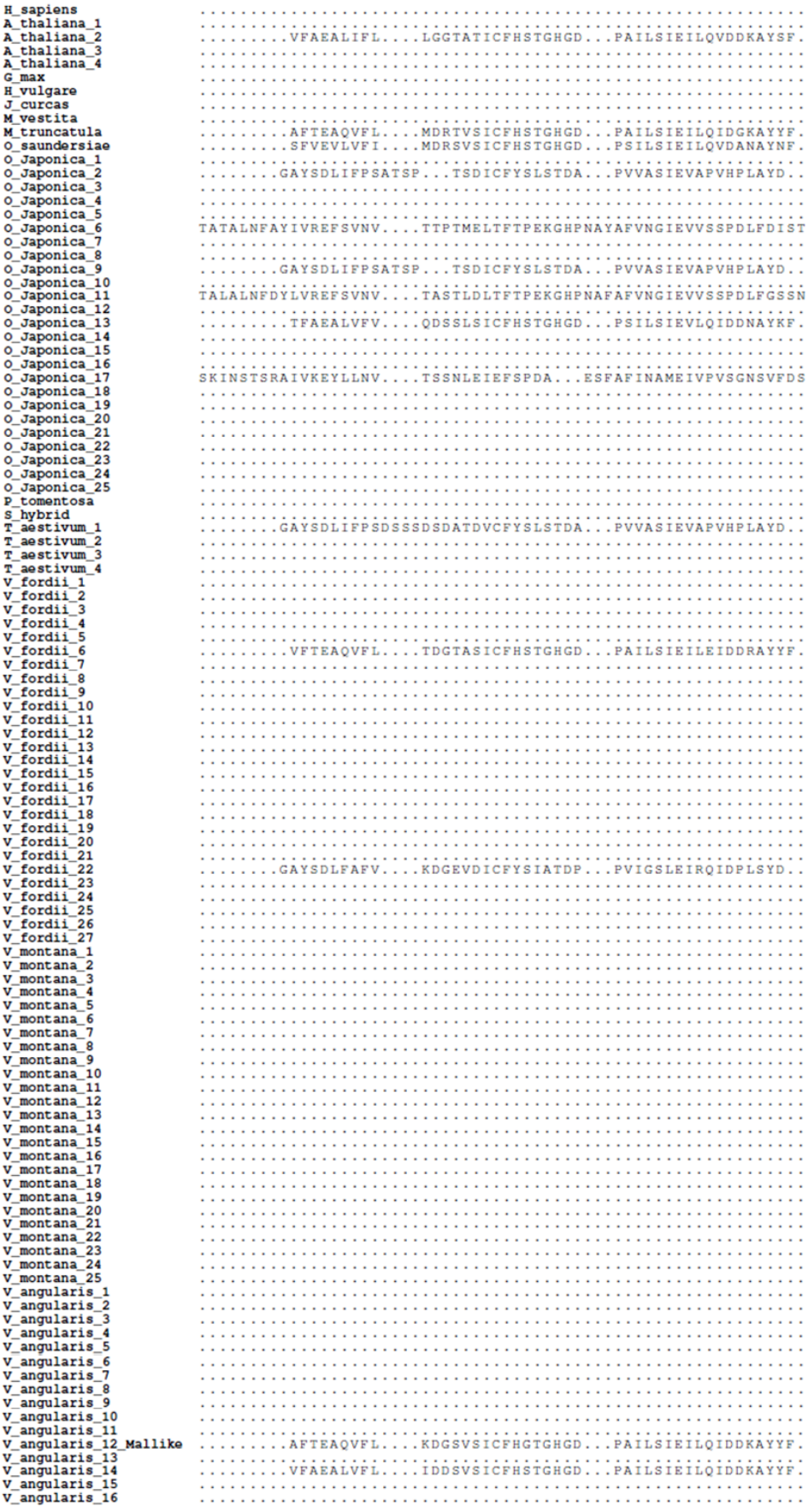
Index information 1:
Luria Bertani (LB) medium culture used for the protocol of expression
Luria Bertani medium culture (1L):
-10g Tryptone (Sigma-Aldrich®);
-10g NaCl ()
-5g Yeast (NZYTech®)
-Required volume of distilled water
Autoclave for 20 minutes at 120ºC

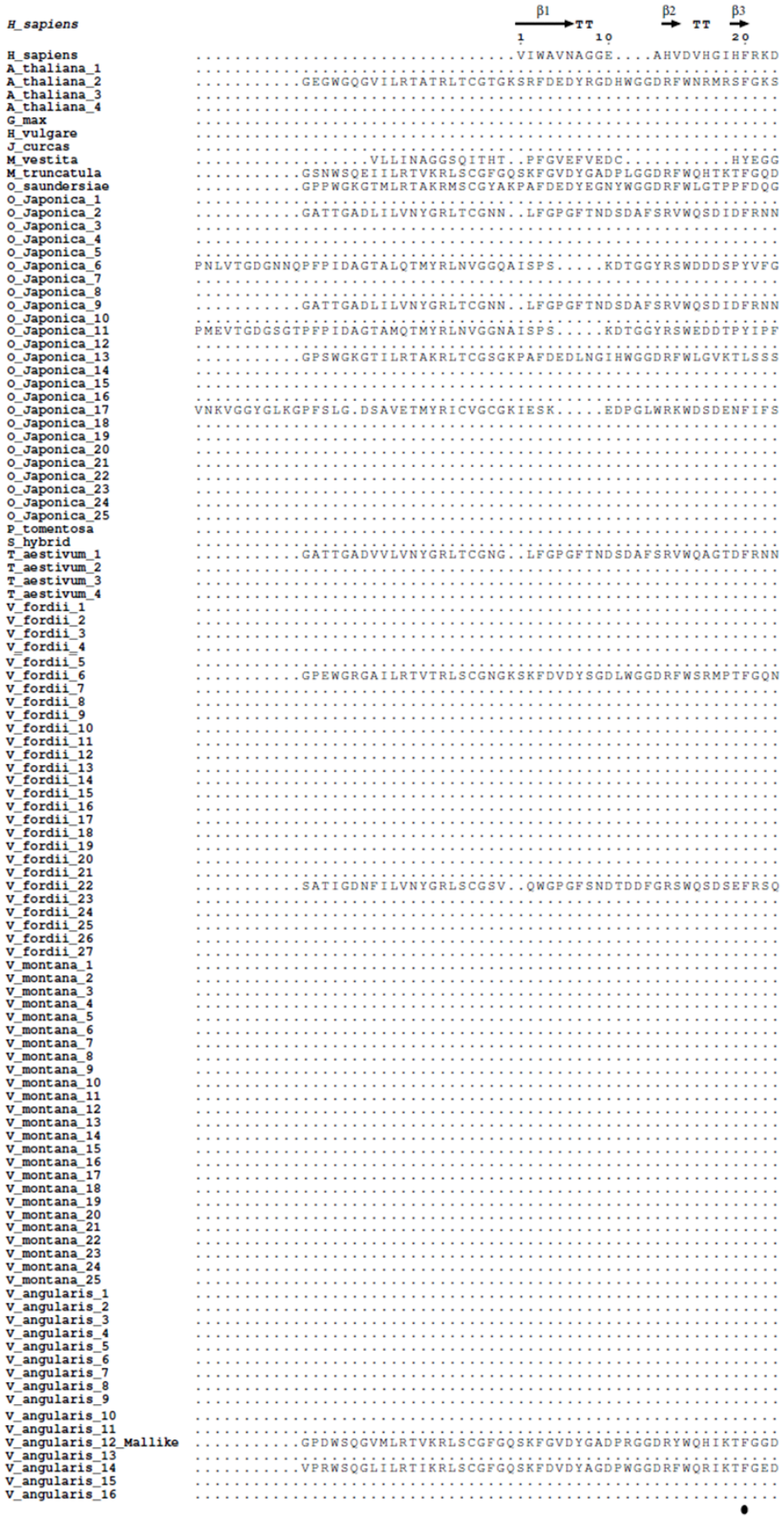

H_sapiens
H_sapiens
A_thaliana_1
A_thaliana_2
A_thaliana_3
A_thaliana_4
G_max
H_vulgare
J_curcas
M_vestita
M_truncatula
O_saundersiae
O_Japonica_1
O_Japonica_2
O_Japonica_3
O_Japonica_4
O_Japonica_5
O_Japonica_6
O_Japonica_7
O_Japonica_8
O_Japonica_9
O_Japonica_10
O_Japonica_11
O_Japonica_12
O_Japonica_13
O_Japonica_14
O_Japonica_15
O_Japonica_16
O_Japonica_17
O_Japonica_18
O_Japonica_19
O_Japonica_20
O_Japonica_21
O_Japonica_22
O_Japonica_23
O_Japonica_24
O_Japonica_25
P_tomentosa
S_hybrid
T_aestivum_1
T_aestivum_2
T_aestivum_3
T_aestivum_4
V_fordii_1
V_fordii_2
V_fordii_3
V_fordii_4
V_fordii_5
V_fordii_6
V_fordii_7
V_fordii_8
V_fordii_9
V_fordii_10
V_fordii_11
V_fordii_12
V_fordii_13
V_fordii_14
V_fordii_15
V_fordii_16
V_fordii_17
V_fordii_18
V_fordii_19
V_fordii_20
V_fordii_21
V_fordii_22
V_fordii_23
V_fordii_24
V_fordii_25
V_fordii_26
V_fordii_27
V_montana_1
V_montana_2
V_montana_3
V_montana_4
V_montana_5
V_montana_6
V_montana_7
V_montana_8
V_montana_9
V_montana_10
V_montana_11
V_montana_12
V_montana_13
V_montana_14
V_montana_15
V_montana_16
V_montana_17
V_montana_18
V_montana_19
V_montana_20
V_montana_21
V_montana_22
V_montana_23
V_montana_24
V_montana_25
V_angularis_1
V_angularis_2
V_angularis_3
V_angularis_4
V_angularis_5
V_angularis_6
V_angularis_7
V_angularis_8
V_angularis_9
V_angularis_10
V_angularis_11
V_angularis_12_Mallike
V_angularis_13
V_angularis_14
V_angularis_15
V_angularis_16

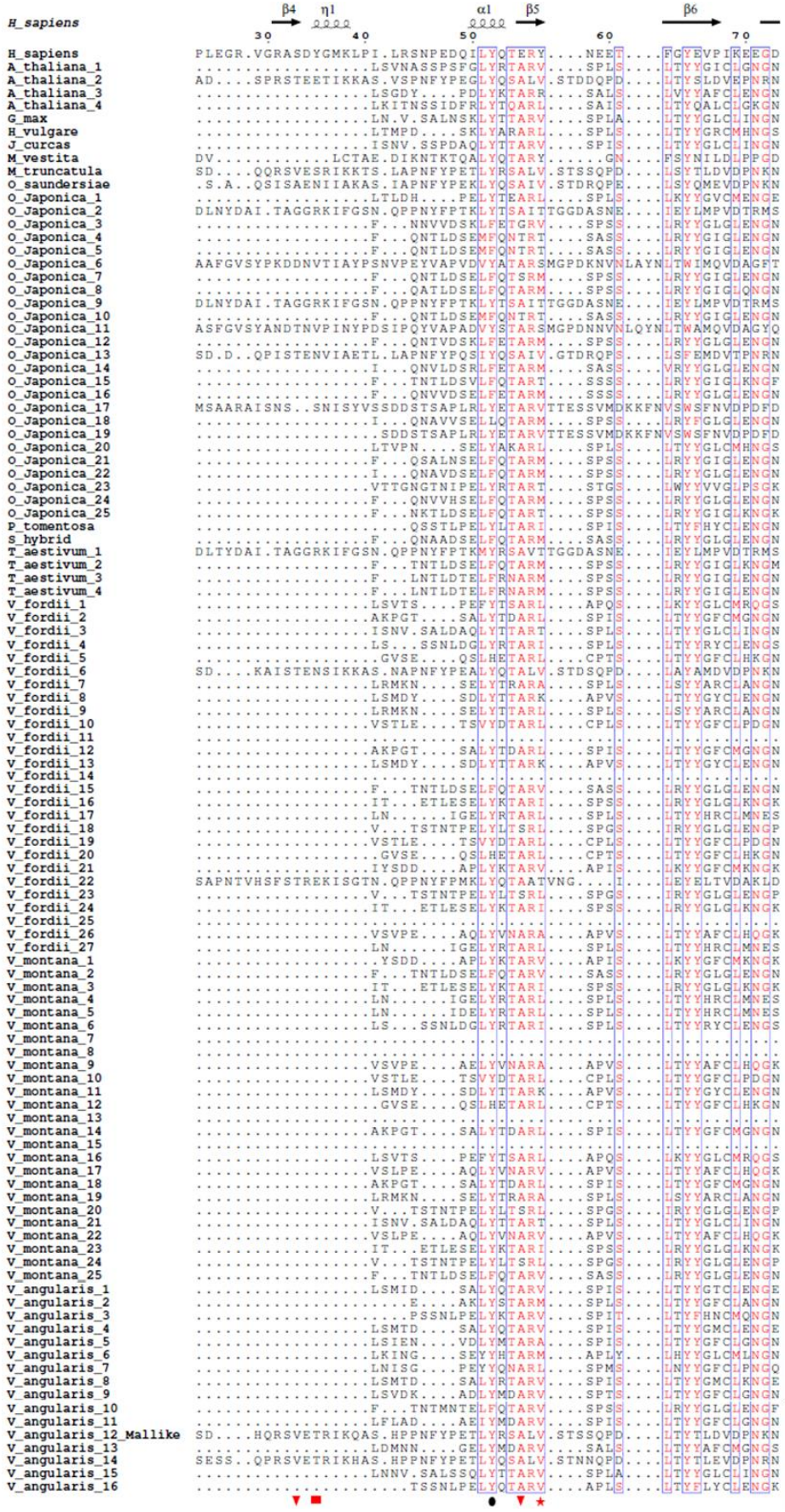

H_sapiens
H_sapiens
A_thaliana_1
A_thaliana_2
A_thaliana_3
A_thaliana_4
G_max
H_vulgare
J_curcas
M_vestita
M_truncatula
O_saundersiae
O_Japonica_1
O_Japonica_2
O_Japonica_3
O_Japonica_4
O_Japonica_5
O_Japonica_6
O_Japonica_7
O_Japonica_8
O_Japonica_9
O_Japonica_10
O_Japonica_11
O_Japonica_12
O_Japonica_13
O_Japonica_14
O_Japonica_15
O_Japonica_16
O_Japonica_17
O_Japonica_18
O_Japonica_19
O_Japonica_20
O_Japonica_21
O_Japonica_22
O_Japonica_23
O_Japonica_24
O_Japonica_25
P_tomentosa
S_hybrid
T_aestivum_1
T_aestivum_2
T_aestivum_3
T_aestivum_4
V_fordii_1
V_fordii_2
V_fordii_3
V_fordii_4
V_fordii_5
V_fordii_6
V_fordii_7
V_fordii_8
V_fordii_9
V_fordii_10
V_fordii_11
V_fordii_12
V_fordii_13
V_fordii_14
V_fordii_15
V_fordii_16
V_fordii_17
V_fordii_18
V_fordii_19
V_fordii_20
V_fordii_21
V_fordii_22
V_fordii_23
V_fordii_24
V_fordii_25
V_fordii_26
V_fordii_27
V_montana_1
V_montana_2
V_montana_3
V_montana_4
V_montana_5
V_montana_6
V_montana_7
V_montana_8
V_montana_9
V_montana_10
V_montana_11
V_montana_12
V_montana_13
V_montana_14
V_montana_15
V_montana_16
V_montana_17
V_montana_18
V_montana_19
V_montana_20
V_montana_21
V_montana_22
V_montana_23
V_montana_24
V_montana_25
V_angularis_1
V_angularis_2
V_angularis_3
V_angularis_4
V_angularis_5
V_angularis_6
V_angularis_7
V_angularis_8
V_angularis_9
V_angularis_10
V_angularis_11
V_angularis_12_Mallike
V_angularis_13
V_angularis_14
V_angularis_15
V_angularis_16

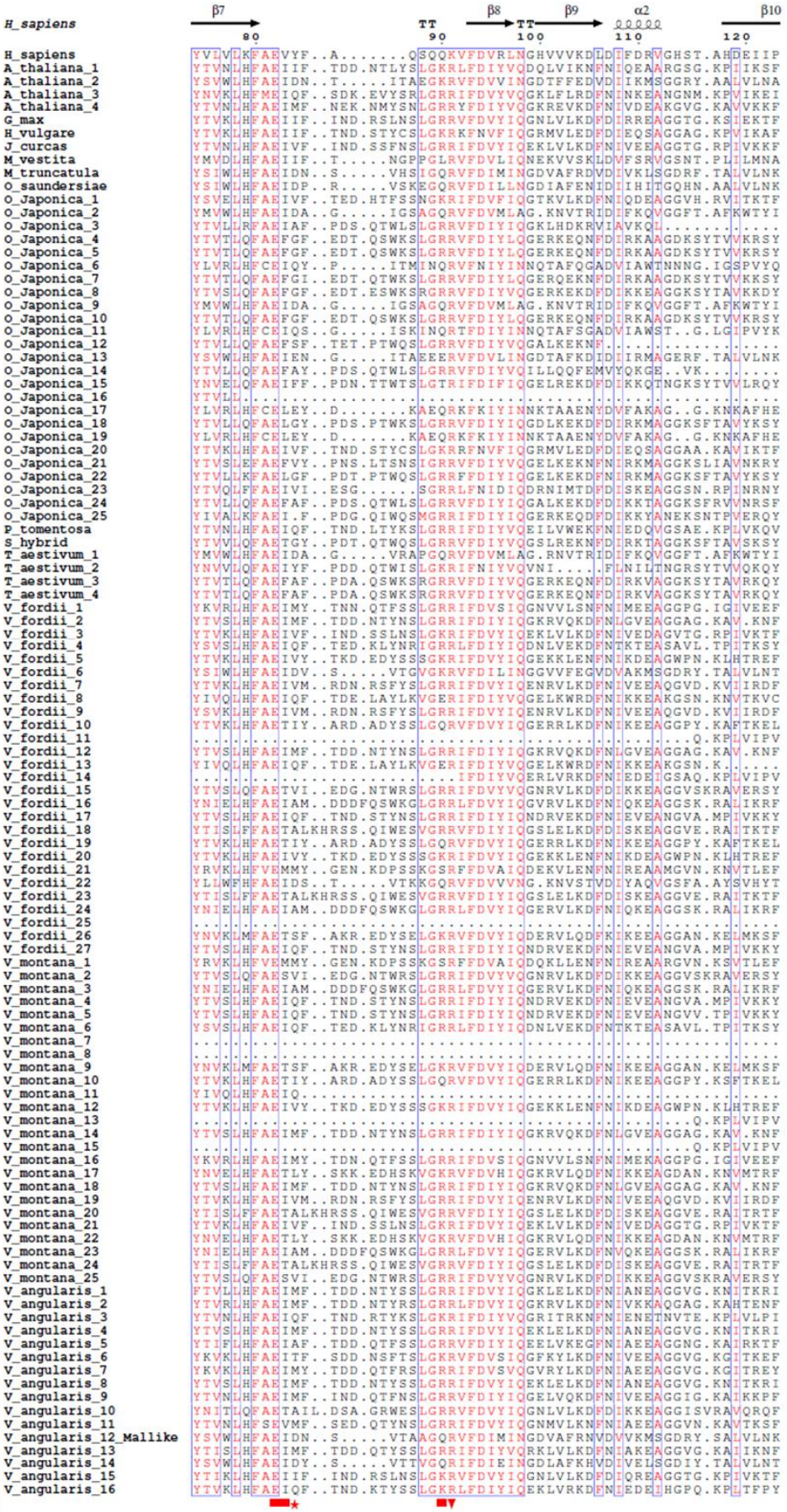
H_sapiens

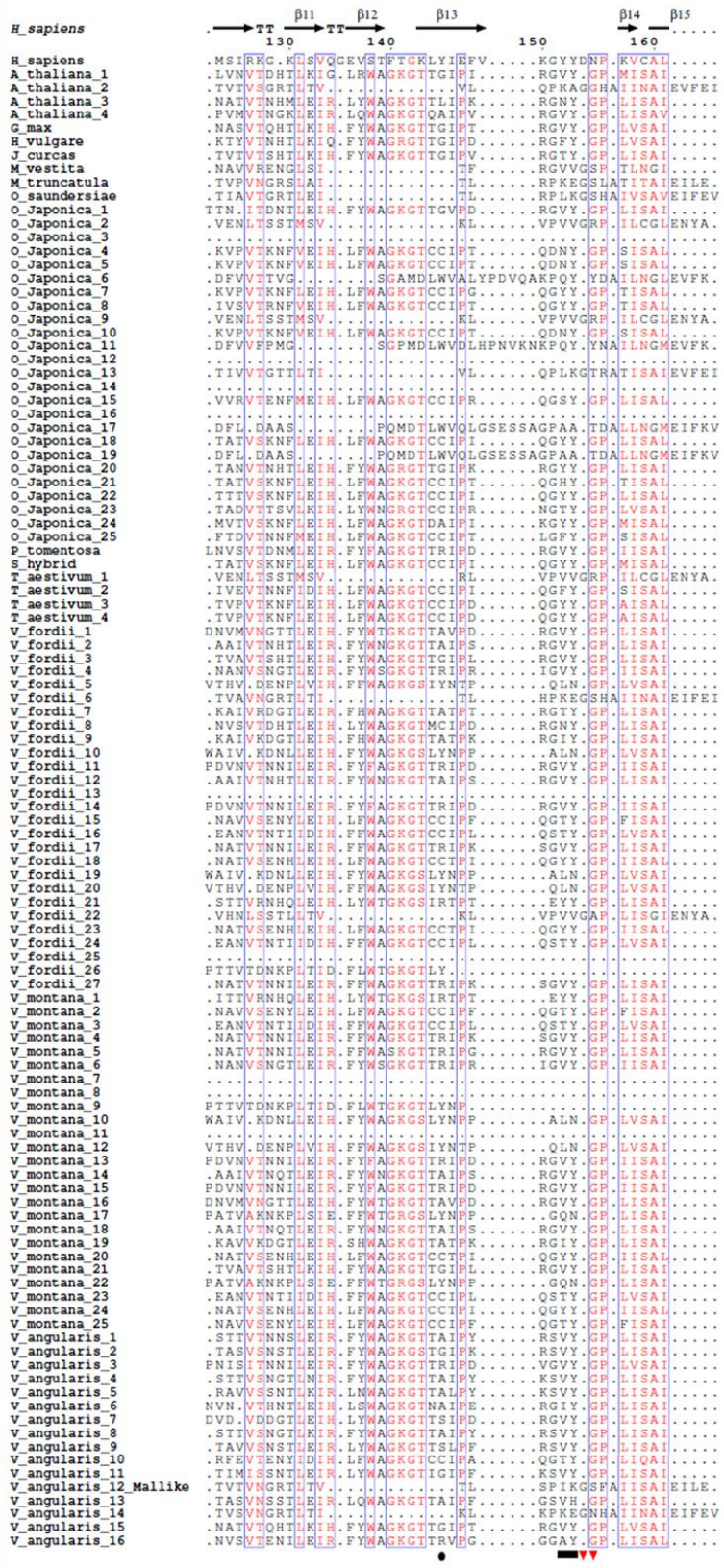
```
H_sapiens                .....................................................................
A_thaliana_1             .....................................................................
A_thaliana_2             NIVRVPKGHYSVRIFFGLVD....QPSFDKEPLFDISIEGTQISSLKSGWSSQDDQ....
A_thaliana_3             .....................................................................
A_thaliana_4             .....................................................................
G_max                    .....................................................................
H_vulgare                .....................................................................
J_curcas                 .....................................................................
M_vestita                .....................................................................
M_truncatula             ..........................NEPLFDISIQGTQIYTLKPGWTTQDDQ....
O_saundersiae            NINNVPNGHYAVRVFFALIA....DPNLENEPVFDISVEGTLIYSLMPGWSNSEDQ....
O_Japonica_1             .....................................................................
O_Japonica_2             S.LPLPPGRYYLRLFSVYDN....YDSKLRTPSFDVSAAATLVLSFRSPWPEPASRY...
O_Japonica_3             .....................................................................
O_Japonica_4             .....................................................................
O_Japonica_5             .....................................................................
O_Japonica_6             SFPL.GAGRKFLRLHFY..PANY.SSRDAADARFSVSVPAANVT.LLSNFSAY......Q
O_Japonica_7             .....................................................................
O_Japonica_8             .....................................................................
O_Japonica_9             S.LPLPPGRYYLRLFSVYDN....YDSKLRTPSFDVSAAATLVLSFRSPWPEPASRY...
O_Japonica_10            .....................................................................
O_Japonica_11            SFPV.GAGRKFLRLHFY..PANY.SNRNAADALFSVSIPDPNIT.LLSNFSAY......Q
O_Japonica_12            .....................................................................
O_Japonica_13            YINNVPNGHYQVRLFFALVA....DPNLDSEPIFDVSVEGTLFSSLLLGWSSEDEK....
O_Japonica_14            .....................................................................
O_Japonica_15            .....................................................................
O_Japonica_16            .....................................................................
O_Japonica_17            KFIVAAAGSYFLRLHFSQLPTNF.STKE...SLFDVSA..NGLK.LVSKFNVPAEIYLRN
O_Japonica_18            .....................................................................
O_Japonica_19            .....................................................................
O_Japonica_20            .....................................................................
O_Japonica_21            .....................................................................
O_Japonica_22            .....................................................................
O_Japonica_23            .....................................................................
O_Japonica_24            .....................................................................
O_Japonica_25            .....................................................................
P_tomentosa              .....................................................................
S_hybrid                 .....................................................................
T_aestivum_1             S.LPLAPGRYYLRVFSVYDN....YNSKLRSPSFDVSAAATLVLSFRSPWPETAARY...
T_aestivum_2             .....................................................................
T_aestivum_3             .....................................................................
T_aestivum_4             .....................................................................
V_fordii_1               .....................................................................
V_fordii_2               .....................................................................
V_fordii_3               .....................................................................
V_fordii_4               .....................................................................
V_fordii_5               .....................................................................
V_fordii_6               NIKRVPKGHYSVRVFFGLVAHHDPDPNFDNEPLFDFSVEGTQFYSLQSGWSSHDDQ....
V_fordii_7               .....................................................................
V_fordii_8               .....................................................................
V_fordii_9               .....................................................................
V_fordii_10              .....................................................................
V_fordii_11              .....................................................................
V_fordii_12              .....................................................................
V_fordii_13              .....................................................................
V_fordii_14              .....................................................................
V_fordii_15              .....................................................................
V_fordii_16              .....................................................................
V_fordii_17              .....................................................................
V_fordii_18              .....................................................................
V_fordii_19              .....................................................................
V_fordii_20              .....................................................................
V_fordii_21              .....................................................................
V_fordii_22              I.LPLPNGRYYIRTFTVYDN....YDGKSHSPSFDASVEGTLVFSWRSPWPENLARD...
V_fordii_23              .....................................................................
V_fordii_24              .....................................................................
V_fordii_25              FHYCFKNGKY..........................................................
V_fordii_26              .....................................................................
V_fordii_27              .....................................................................
V_montana_1              .....................................................................
V_montana_2              .....................................................................
V_montana_3              .....................................................................
V_montana_4              .....................................................................
V_montana_5              .....................................................................
V_montana_6              .....................................................................
V_montana_7              FHYCFKNGKY..........................................................
V_montana_8              FHYCFKNGKY..........................................................
V_montana_9              .....................................................................
V_montana_10             .....................................................................
V_montana_11             .....................................................................
V_montana_12             .....................................................................
V_montana_13             .....................................................................
V_montana_14             .....................................................................
V_montana_15             .....................................................................
V_montana_16             .....................................................................
V_montana_17             .....................................................................
V_montana_18             .....................................................................
V_montana_19             .....................................................................
V_montana_20             .....................................................................
V_montana_21             .....................................................................
V_montana_22             .....................................................................
V_montana_23             .....................................................................
V_montana_24             .....................................................................
V_montana_25             .....................................................................
V_angularis_1            .....................................................................
V_angularis_2            .....................................................................
V_angularis_3            .....................................................................
V_angularis_4            .....................................................................
V_angularis_5            .....................................................................
V_angularis_6            .....................................................................
V_angularis_7            .....................................................................
V_angularis_8            .....................................................................
V_angularis_9            .....................................................................
V_angularis_10           .....................................................................
V_angularis_11           .....................................................................
V_angularis_12_Mallike   NINRVPKGRYSIRIFFGLVAHTT....VTDEPLFDISIQGTQIYSLKSGWTTQDDQ....
V_angularis_13           .....................................................................
V_angularis_14           VIDKVPKGHYLIRIFFGLVGQSR...GTIIEPLFDISIEGTQIYSLKSGWTHQADQ....
V_angularis_15           .....................................................................
V_angularis_16           .....................................................................
```


H_sapiens
A_thaliana_1
A_thaliana_2
A_thaliana_3
A_thaliana_4
G_max
H_vulgare
J_curcas
M_vestita
M_truncatula
O_saundersiae
O_Japonica_1
O_Japonica_2
O_Japonica_3
O_Japonica_4
O_Japonica_5
O_Japonica_6
O_Japonica_7
O_Japonica_8
O_Japonica_9
O_Japonica_10
O_Japonica_11
O_Japonica_12
O_Japonica_13
O_Japonica_14
O_Japonica_15
O_Japonica_16
O_Japonica_17
O_Japonica_18
O_Japonica_19
O_Japonica_20
O_Japonica_21
O_Japonica_22
O_Japonica_23
O_Japonica_24
O_Japonica_25
P_tomentosa
S_hybrid
T_aestivum_1
T_aestivum_2
T_aestivum_3
T_aestivum_4
V_fordii_1
V_fordii_2
V_fordii_3
V_fordii_4
V_fordii_5
V_fordii_6
V_fordii_7
V_fordii_8
V_fordii_9
V_fordii_10
V_fordii_11
V_fordii_12
V_fordii_13
V_fordii_14
V_fordii_15
V_fordii_16
V_fordii_17
V_fordii_18
V_fordii_19
V_fordii_20
V_fordii_21
V_fordii_22
V_fordii_23
V_fordii_24
V_fordii_25
V_fordii_26
V_fordii_27
V_montana_1
V_montana_2
V_montana_3
V_montana_4
V_montana_5
V_montana_6
V_montana_7
V_montana_8
V_montana_9
V_montana_10
V_montana_11
V_montana_12
V_montana_13
V_montana_14
V_montana_15
V_montana_16
V_montana_17
V_montana_18
V_montana_19
V_montana_20
V_montana_21
V_montana_22
V_montana_23
V_montana_24
V_montana_25
V_angularis_1
V_angularis_2
V_angularis_3
V_angularis_4
V_angularis_5
V_angularis_6
V_angularis_7
V_angularis_8
V_angularis_9
V_angularis_10
V_angularis_11
V_angularis_12_Mallike
V_angularis_13
V_angularis_14
V_angularis_15
V_angularis_16

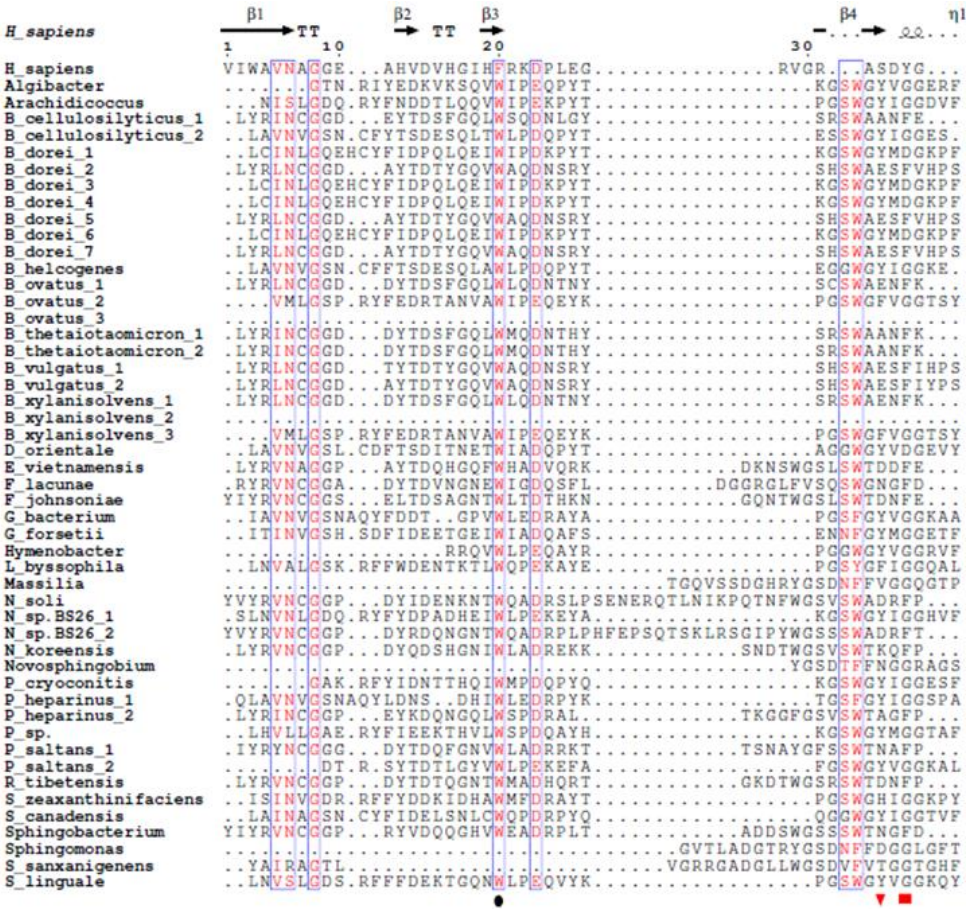

H_sapiens
β1
TT
β2
TT
β3
1
10
20
H_sapiens .....VIWAVNAGGE....AHVDVHGIHFRKD
A_thaliana_1 .....
A_thaliana_2 .....GEGWGQGVILRTATRLTCGTGKSRFDEDYRGDHWGGDRFWNRMRSFGKS
A_thaliana_3 .....
A_thaliana_4 .....
G_max .....
H_vulgare .....
J_curcas .....
M_vestita .....VLLINAGGSQITHT..PFGVEFVEDC.....HYEGG
M_truncatula .....GSNWSQEIILRTVKRLSCGFGQSKFGVDYGADPLGGDRFWQHTKTFGQD
O_saundersiae .....GPPWGKGTMLRTAKRMSCGYAKPAFDEDYEGNYWGGDRFWLGTPPFDQG
O_Japonica_1 .....
O_Japonica_2 .....GATTGADLILVNYGRLTCGNN..LFGPGFTNDSDAFSRVWQSDIDFRNN
O_Japonica_3 .....
O_Japonica_4 .....
O_Japonica_5 .....
O_Japonica_6 PNLVTGDGNNQPFPIDAGTALQTMYRLNVGGQAISPS.....KDTGGYRSWDDDSPYVFG
O_Japonica_7 .....
O_Japonica_8 .....
O_Japonica_9 .....GATTGADLILVNYGRLTCGNN..LFGPGFTNDSDAFSRVWQSDIDFRNN
O_Japonica_10 .....
O_Japonica_11 PMEVTGDGSGTPFPIDAGTAMQTMYRLNVGGNAISPS.....KDTGGYRSWEDDTPYIPF
O_Japonica_12 .....
O_Japonica_13 .....GPSWGKGTILRTAKRLTCGSGKPAFDEDLNGIHWGGDRFWLGVKTLSSS
O_Japonica_14 .....
O_Japonica_15 .....
O_Japonica_16 .....
O_Japonica_17 VNKVGGYGLKGPFSLG.DSAVETMYRICVGCGKIESK.....EDPGLWRKWDSDENFIFS
O_Japonica_18 .....
O_Japonica_19 .....
O_Japonica_20 .....
O_Japonica_21 .....
O_Japonica_22 .....
O_Japonica_23 .....
O_Japonica_24 .....
O_Japonica_25 .....
P_tomentosa .....
S_hybrid .....
T_aestivum_1 .....GATTGADVVLVNYGRLTCGNG..LFGPGFTNDSDAFSRVWQAGTDFRNN
T_aestivum_2 .....
T_aestivum_3 .....
T_aestivum_4 .....
V_fordii_1 .....
V_fordii_2 .....
V_fordii_3 .....
V_fordii_4 .....
V_fordii_5 .....
V_fordii_6 .....GPEWGRGAILRTVTRLSCGNGKSKFDVDYSGDLWGGDRFWSRMPTFGQN
V_fordii_7 .....
V_fordii_8 .....
V_fordii_9 .....
V_fordii_10 .....
V_fordii_11 .....
V_fordii_12 .....
V_fordii_13 .....
V_fordii_14 .....
V_fordii_15 .....
V_fordii_16 .....
V_fordii_17 .....
V_fordii_18 .....
V_fordii_19 .....
V_fordii_20 .....
V_fordii_21 .....
V_fordii_22 .....SATIGDNFILVNYGRLSCGSV..QWGPGFSNDTDDFGRSWQSDSEFRSQ
V_fordii_23 .....
V_fordii_24 .....
V_fordii_25 .....
V_fordii_26 .....
V_fordii_27 .....
V_montana_1 .....
V_montana_2 .....
V_montana_3 .....
V_montana_4 .....
V_montana_5 .....
V_montana_6 .....
V_montana_7 .....
V_montana_8 .....
V_montana_9 .....
V_montana_10 .....
V_montana_11 .....
V_montana_12 .....
V_montana_13 .....
V_montana_14 .....
V_montana_15 .....
V_montana_16 .....
V_montana_17 .....
V_montana_18 .....
V_montana_19 .....
V_montana_20 .....
V_montana_21 .....
V_montana_22 .....
V_montana_23 .....
V_montana_24 .....
V_montana_25 .....
V_angularis_1 .....
V_angularis_2 .....
V_angularis_3 .....
V_angularis_4 .....
V_angularis_5 .....
V_angularis_6 .....
V_angularis_7 .....
V_angularis_8 .....
V_angularis_9 .....
V_angularis_10 .....
V_angularis_11 .....
V_angularis_12_Mallike .....GPDWSQGVMLRTVKRLSCGFGQSKFGVDYGADPRGGDRYWQHIKTFGGD
V_angularis_13 .....
V_angularis_14 .....VPRWSQGLILRTIKRLSCGFGQSKFDVDYAGDPWGGDRFWQRIKTFGED
V_angularis_15 .....
V_angularis_16 .....

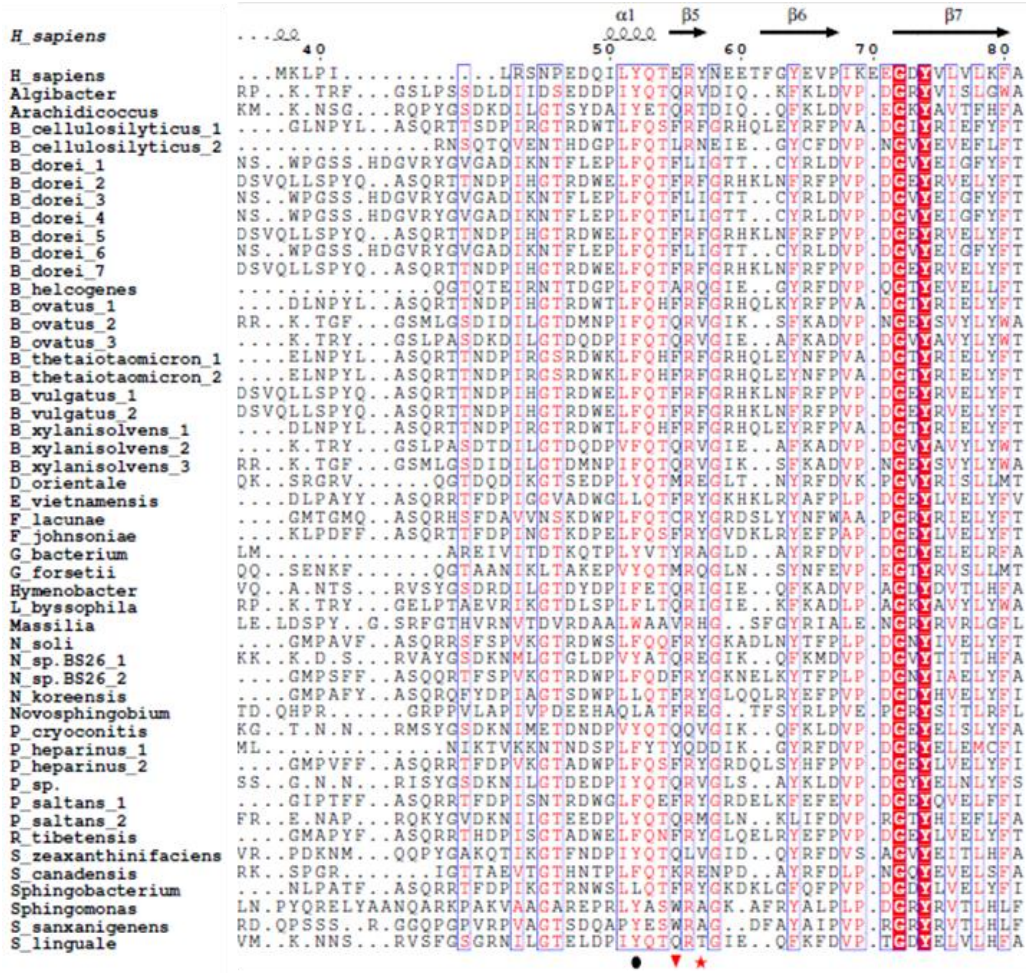

H_sapiens
β4 η1 α1 β5 β6
30 40 50 60 70
H_sapiens
A_thaliana_1
A_thaliana_2
A_thaliana_3
A_thaliana_4
G_max
H_vulgare
J_curcas
M_vestita
M_truncatula
O_saundersiae
O_Japonica_1
O_Japonica_2
O_Japonica_3
O_Japonica_4
O_Japonica_5
O_Japonica_6
O_Japonica_7
O_Japonica_8
O_Japonica_9
O_Japonica_10
O_Japonica_11
O_Japonica_12
O_Japonica_13
O_Japonica_14
O_Japonica_15
O_Japonica_16
O_Japonica_17
O_Japonica_18
O_Japonica_19
O_Japonica_20
O_Japonica_21
O_Japonica_22
O_Japonica_23
O_Japonica_24
O_Japonica_25
P_tomentosa
S_hybrid
T_aestivum_1
T_aestivum_2
T_aestivum_3
T_aestivum_4
V_fordii_1
V_fordii_2
V_fordii_3
V_fordii_4
V_fordii_5
V_fordii_6
V_fordii_7
V_fordii_8
V_fordii_9
V_fordii_10
V_fordii_11
V_fordii_12
V_fordii_13
V_fordii_14
V_fordii_15
V_fordii_16
V_fordii_17
V_fordii_18
V_fordii_19
V_fordii_20
V_fordii_21
V_fordii_22
V_fordii_23
V_fordii_24
V_fordii_25
V_fordii_26
V_fordii_27
V_montana_1
V_montana_2
V_montana_3
V_montana_4
V_montana_5
V_montana_6
V_montana_7
V_montana_8
V_montana_9
V_montana_10
V_montana_11
V_montana_12
V_montana_13
V_montana_14
V_montana_15
V_montana_16
V_montana_17
V_montana_18
V_montana_19
V_montana_20
V_montana_21
V_montana_22
V_montana_23
V_montana_24
V_montana_25
V_angularis_1
V_angularis_2
V_angularis_3
V_angularis_4
V_angularis_5
V_angularis_6
V_angularis_7
V_angularis_8
V_angularis_9
V_angularis_10
V_angularis_11
V_angularis_12_Mallike
V_angularis_13
V_angularis_14
V_angularis_15
V_angularis_16

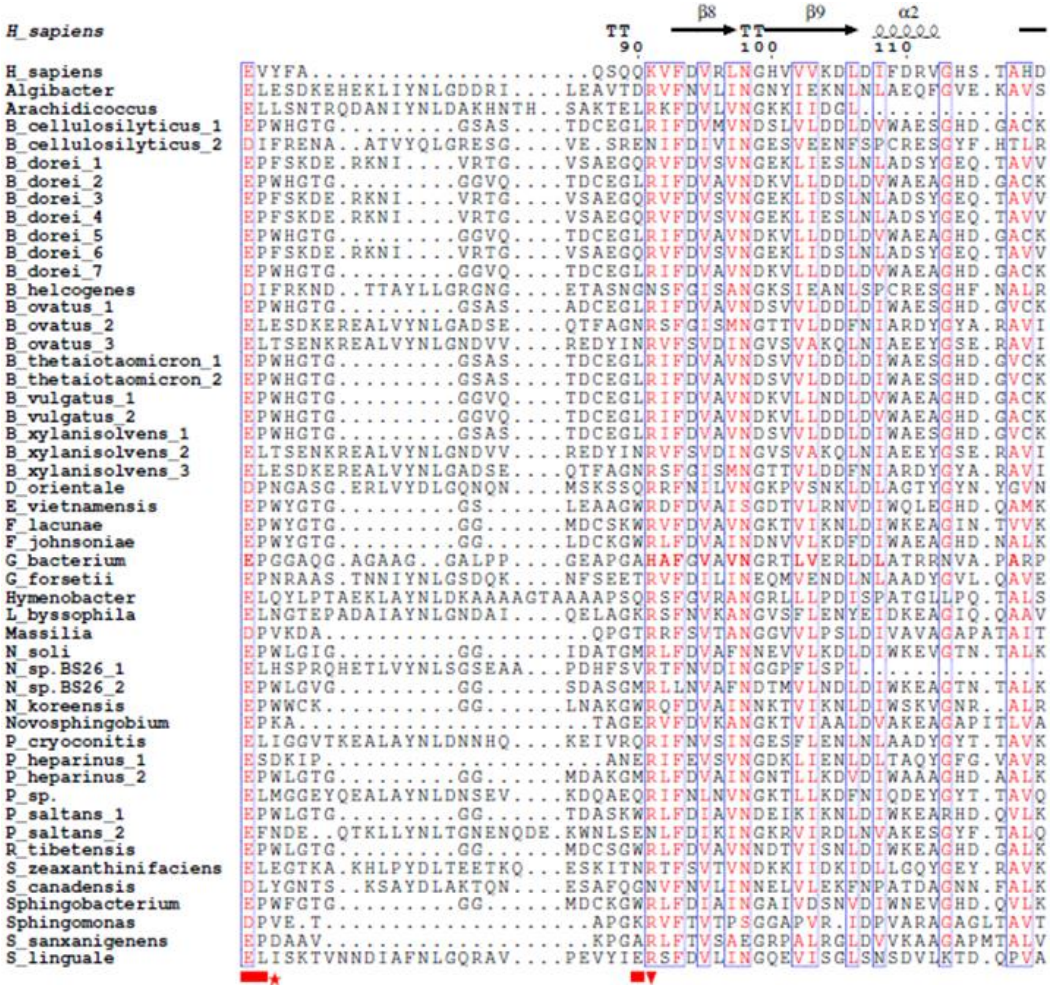

H_sapiens
β7
β8
β9
α2
β10
TT
TT
80
90
100
110
120
H_sapiens
A_thaliana_1
A_thaliana_2
A_thaliana_3
A_thaliana_4
G_max
H_vulgare
J_curcas
M_vestita
M_truncatula
O_saundersiae
O_Japonica_1
O_Japonica_2
O_Japonica_3
O_Japonica_4
O_Japonica_5
O_Japonica_6
O_Japonica_7
O_Japonica_8
O_Japonica_9
O_Japonica_10
O_Japonica_11
O_Japonica_12
O_Japonica_13
O_Japonica_14
O_Japonica_15
O_Japonica_16
O_Japonica_17
O_Japonica_18
O_Japonica_19
O_Japonica_20
O_Japonica_21
O_Japonica_22
O_Japonica_23
O_Japonica_24
O_Japonica_25
P_tomentosa
S_hybrid
T_aestivum_1
T_aestivum_2
T_aestivum_3
T_aestivum_4
V_fordii_1
V_fordii_2
V_fordii_3
V_fordii_4
V_fordii_5
V_fordii_6
V_fordii_7
V_fordii_8
V_fordii_9
V_fordii_10
V_fordii_11
V_fordii_12
V_fordii_13
V_fordii_14
V_fordii_15
V_fordii_16
V_fordii_17
V_fordii_18
V_fordii_19
V_fordii_20
V_fordii_21
V_fordii_22
V_fordii_23
V_fordii_24
V_fordii_25
V_fordii_26
V_fordii_27
V_montana_1
V_montana_2
V_montana_3
V_montana_4
V_montana_5
V_montana_6
V_montana_7
V_montana_8
V_montana_9
V_montana_10
V_montana_11
V_montana_12
V_montana_13
V_montana_14
V_montana_15
V_montana_16
V_montana_17
V_montana_18
V_montana_19
V_montana_20
V_montana_21
V_montana_22
V_montana_23
V_montana_24
V_montana_25
V_angularis_1
V_angularis_2
V_angularis_3
V_angularis_4
V_angularis_5
V_angularis_6
V_angularis_7
V_angularis_8
V_angularis_9
V_angularis_10
V_angularis_11
V_angularis_12_Mallike
V_angularis_13
V_angularis_14
V_angularis_15
V_angularis_16

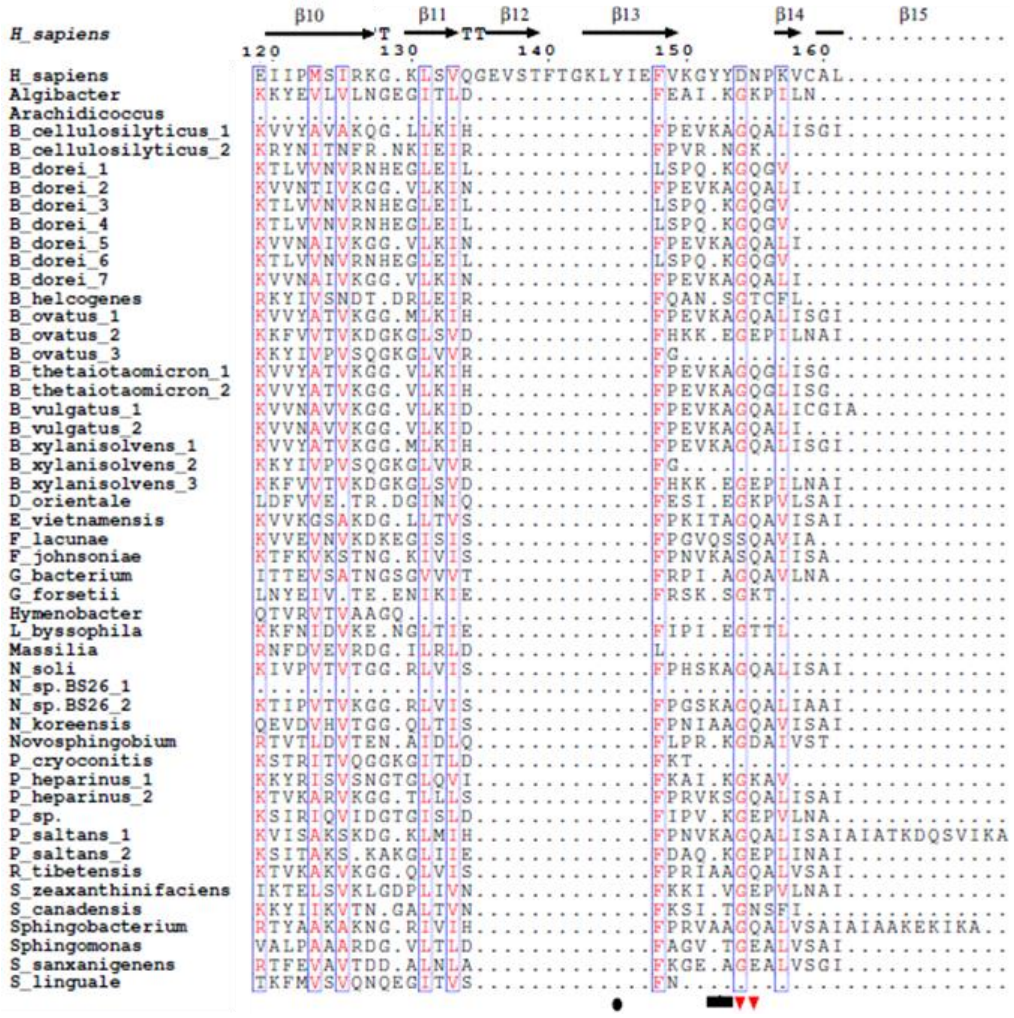


Index Figure 1- Alignment of each malectin-like amino acid sequence in plants compared to the human malectin, using ClustalOmega [Sievers, *et al*, 2011] and ESPript [Robert, *et al*, 2014] programs. The amino acids in red are conserved. The red symbols above the alignment represents the binding sites residues; Red triangles-by direct hydrogen bonds; Red squares-by hydrogen bonds mediated by water; red stars- by π/CH interactions. The black symbols mark the carbohydrate-interacting residues from malectin putative binding site.

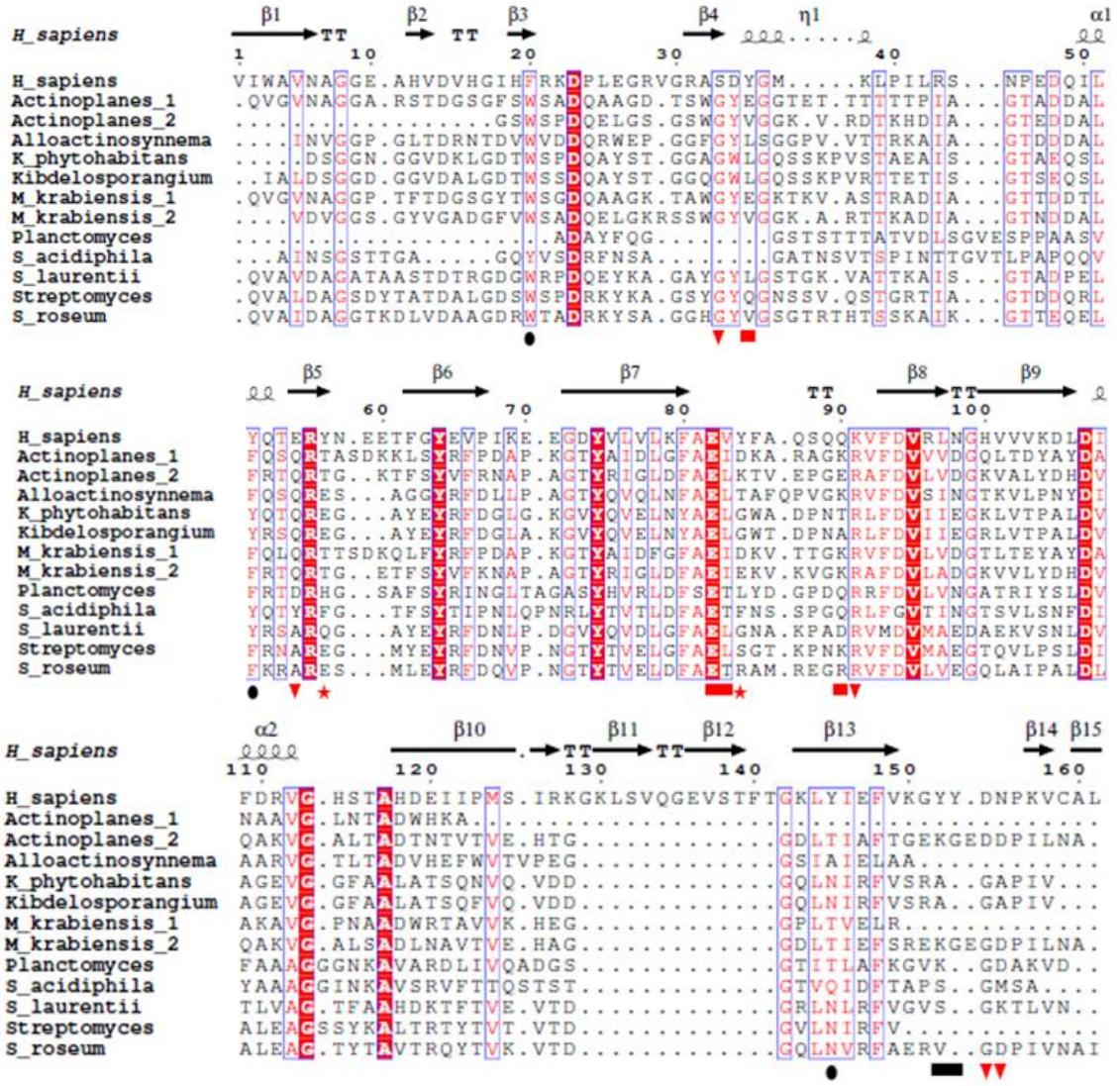

β1
β2
β3
β4
η1
H_sapiens
H_sapiens
Algibacter
Arachidicoccus
B_cellulosilyticus_1
B_cellulosilyticus_2
B_dorei_1
B_dorei_2
B_dorei_3
B_dorei_4
B_dorei_5
B_dorei_6
B_dorei_7
B_helcogenes
B_ovatus_1
B_ovatus_2
B_ovatus_3
B_thetaiotaomicron_1
B_thetaiotaomicron_2
B_vulgatus_1
B_vulgatus_2
B_xylanisolvens_1
B_xylanisolvens_2
B_xylanisolvens_3
D_orientale
E_vietnamensis
F_lacunae
F_johnsoniae
G_bacterium
G_forsetii
Hymenobacter
L_byssophila
Massilia
N_soli
N_sp.BS26_1
N_sp.BS26_2
N_koreensis
Novosphingobium
P_cryoconitis
P_heparinus_1
P_heparinus_2
P_sp.
P_saltans_1
P_saltans_2
R_tibetensis
S_zeaxanthinifaciens
S_canadensis
Sphingobacterium
Sphingomonas
S_sanxanigenens
S_linguale


α1
β5
β6
β7
H_sapiens
40
50
60
70
80
H_sapiens
Algibacter
Arachidicoccus
B_cellulosilyticus_1
B_cellulosilyticus_2
B_dorei_1
B_dorei_2
B_dorei_3
B_dorei_4
B_dorei_5
B_dorei_6
B_dorei_7
B_helcogenes
B_ovatus_1
B_ovatus_2
B_ovatus_3
B_thetaiotaomicron_1
B_thetaiotaomicron_2
B_vulgatus_1
B_vulgatus_2
B_xylanisolvens_1
B_xylanisolvens_2
B_xylanisolvens_3
D_orientale
E_vietnamensis
F_lacunae
F_johnsoniae
G_bacterium
G_forsetii
Hymenobacter
L_byssophila
Massilia
N_soli
N_sp.BS26_1
N_sp.BS26_2
N_koreensis
Novosphingobium
P_cryoconitis
P_heparinus_1
P_heparinus_2
p_sp.
P_saltans_1
P_saltans_2
R_tibetensis
S_zeaxanthinifaciens
S_canadensis
Sphingobacterium
Sphingomonas
S_sanxanigenens
S_linguale

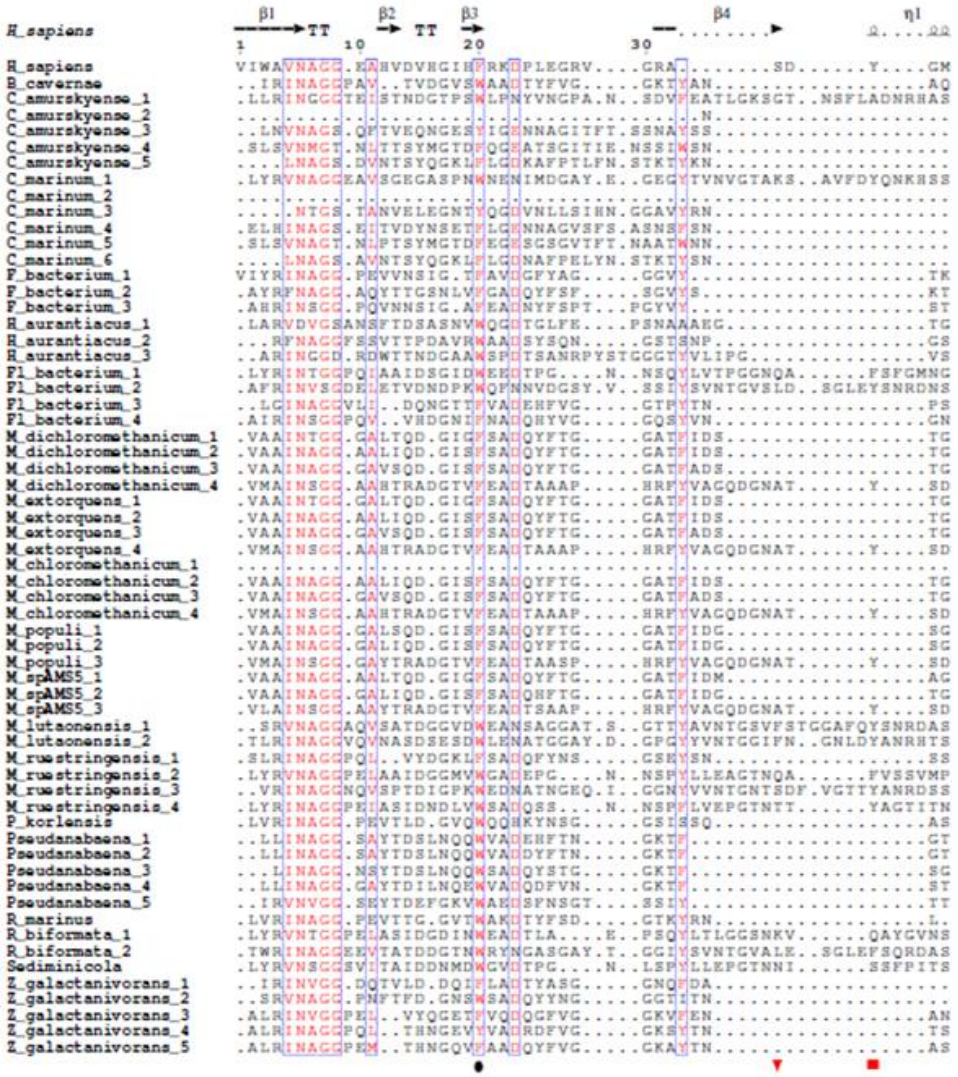

H_sapiens
β8
β9
α2
TT
TT
90
100
110
H_sapiens
Algibacter
Arachidicoccus
B_cellulosilyticus_1
B_cellulosilyticus_2
B_dorei_1
B_dorei_2
B_dorei_3
B_dorei_4
B_dorei_5
B_dorei_6
B_dorei_7
B_helcogenes
B_ovatus_1
B_ovatus_2
B_ovatus_3
B_thetaiotaomicron_1
B_thetaiotaomicron_2
B_vulgatus_1
B_vulgatus_2
B_xylanisolvens_1
B_xylanisolvens_2
B_xylanisolvens_3
D_orientale
E_vietnamensis
F_lacunae
F_johnsoniae
G_bacterium
G_forsetii
Hymenobacter
L_byssophila
Massilia
N_soli
N_sp.BS26_1
N_sp.BS26_2
N_koreensis
Novosphingobium
P_cryoconitis
P_heparinus_1
P_heparinus_2
P_sp.
P_saltans_1
P_saltans_2
R_tibetensis
S_zeaxanthinifaciens
S_canadensis
Sphingobacterium
Sphingomonas
S_sanxanigenens
S_linguale

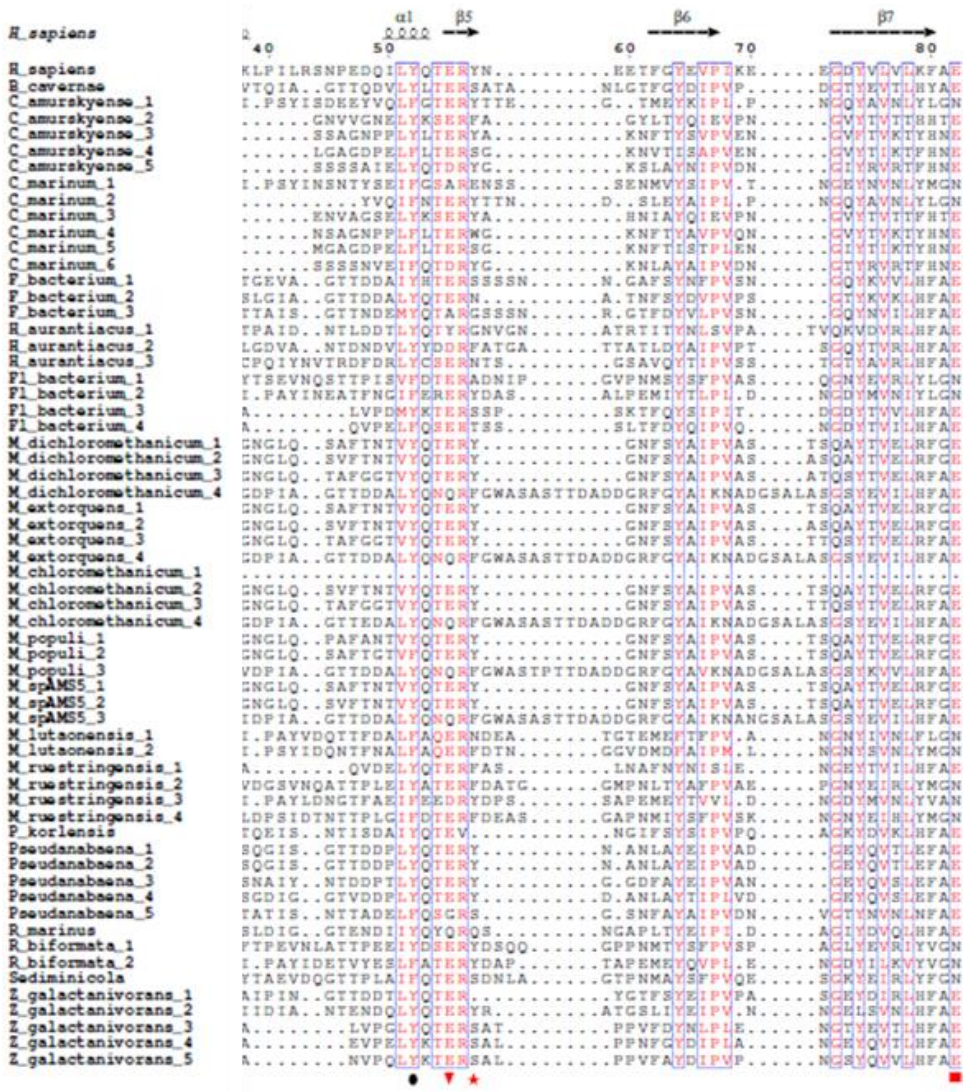


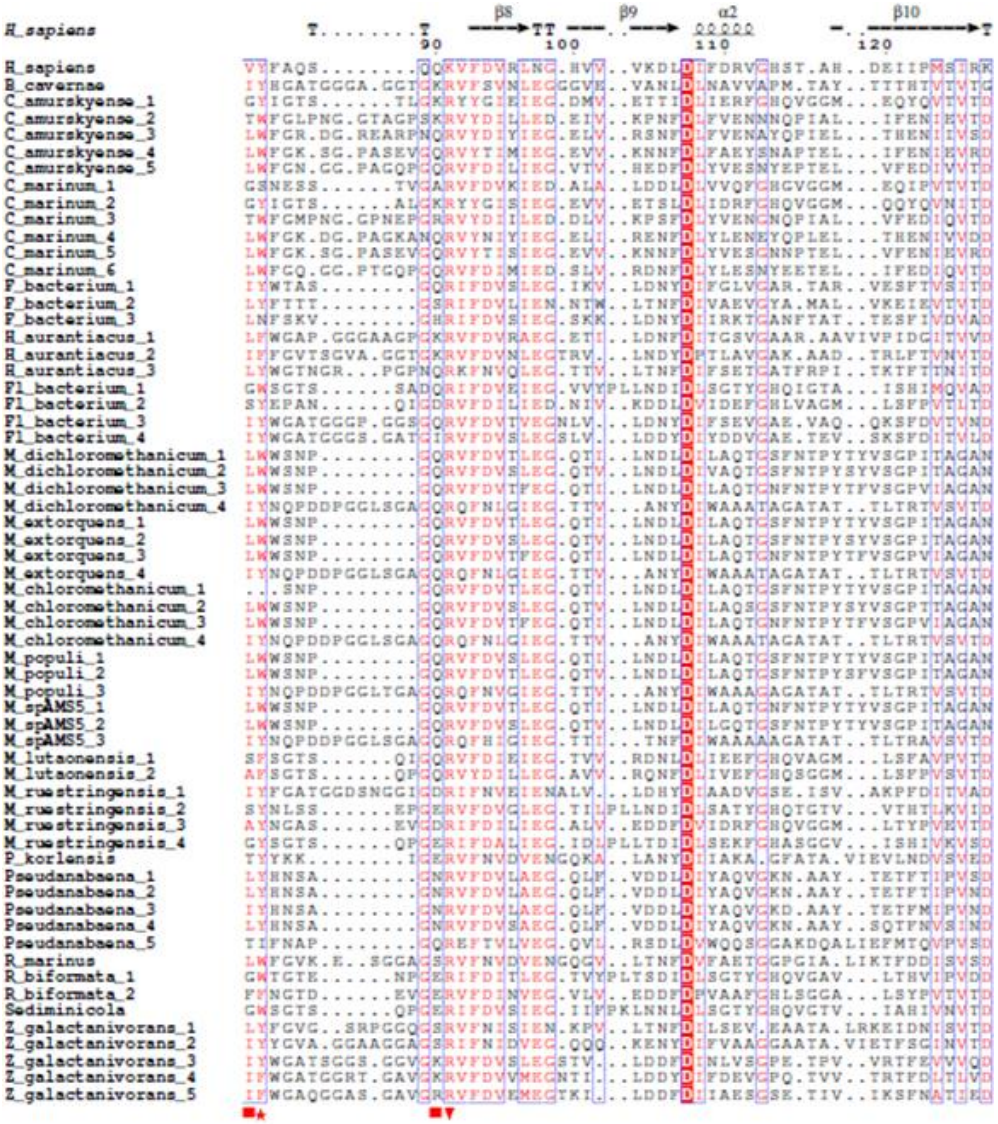
Index Figure 2- Alignment of CBM57 modules associated with glycoside hydrolase family 2 compared to the human malectin, using ClustalOmega [Sievers, *et al*, 2011] and ESPript [Robert, *et al*, 2014] programs. The amino acids in red are conserved. The red symbols above the alignment represents the binding sites residues; Red triangles-by direct hydrogen bonds; Red squares-by hydrogen bonds mediated by water; red stars- by π/CH interactions. The black symbols mark the carbohydrate-interacting residues from malectin putative binding site.

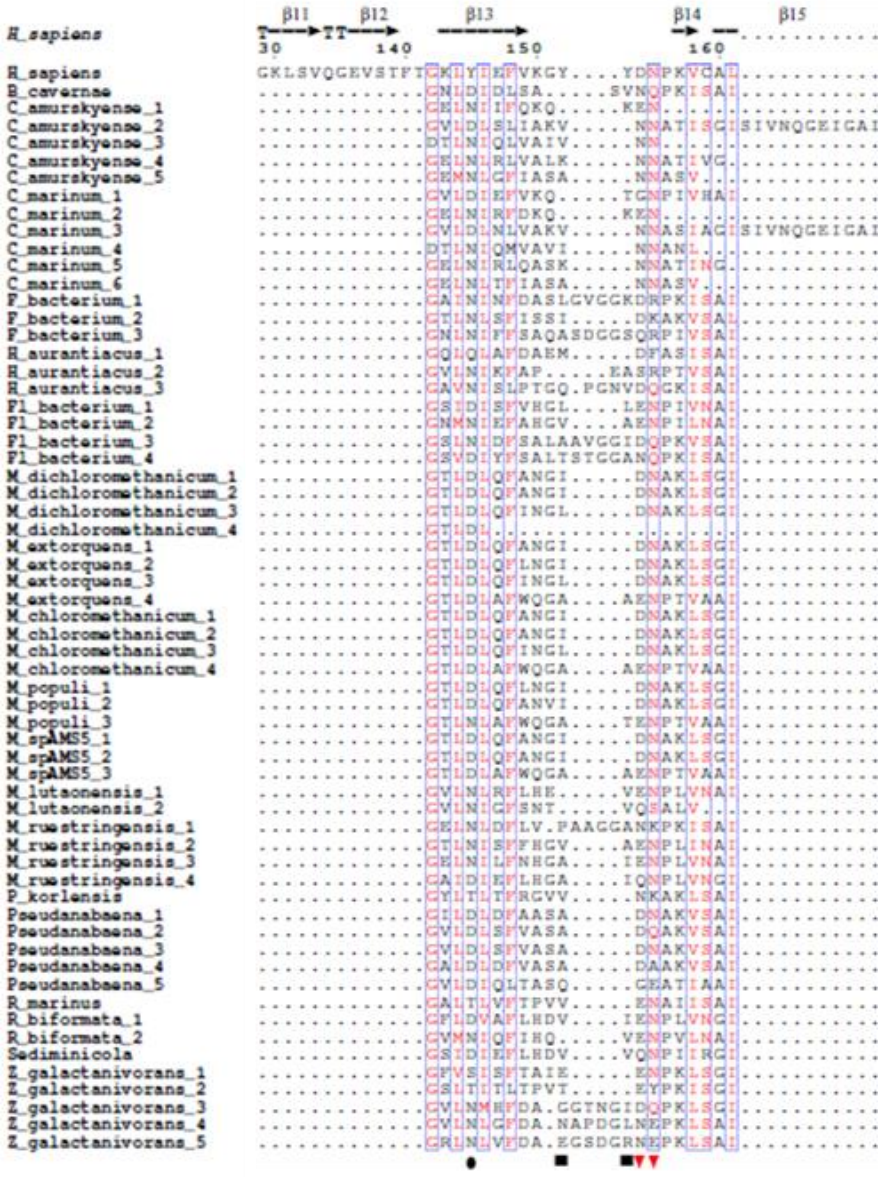


Index Figure 3- Alignment of CBM57 modules associated with peptidase S8/S53 compared to the human malectin, using ClustalOmega [Sievers, *et al*, 2011] and ESPript [Robert, *et al*, 2014] programs. The amino acids in red are conserved. The red symbols above the alignment represents the binding sites residues; Red triangles-by direct hydrogen bonds; Red squares-by hydrogen bonds mediated by water; red stars- by π/CH interactions. The black symbols mark the carbohydrate-interacting residues from malectin putative binding site.

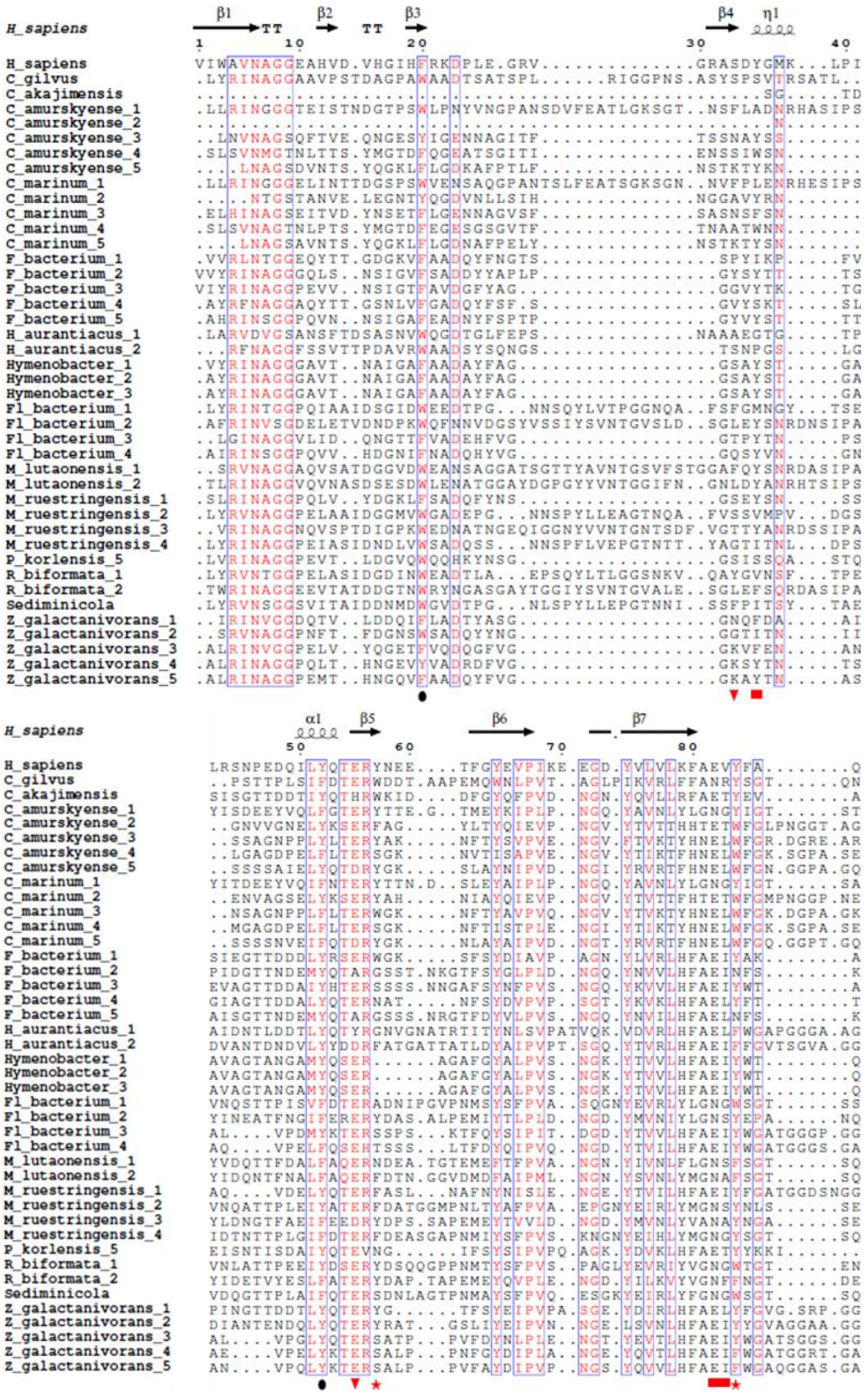


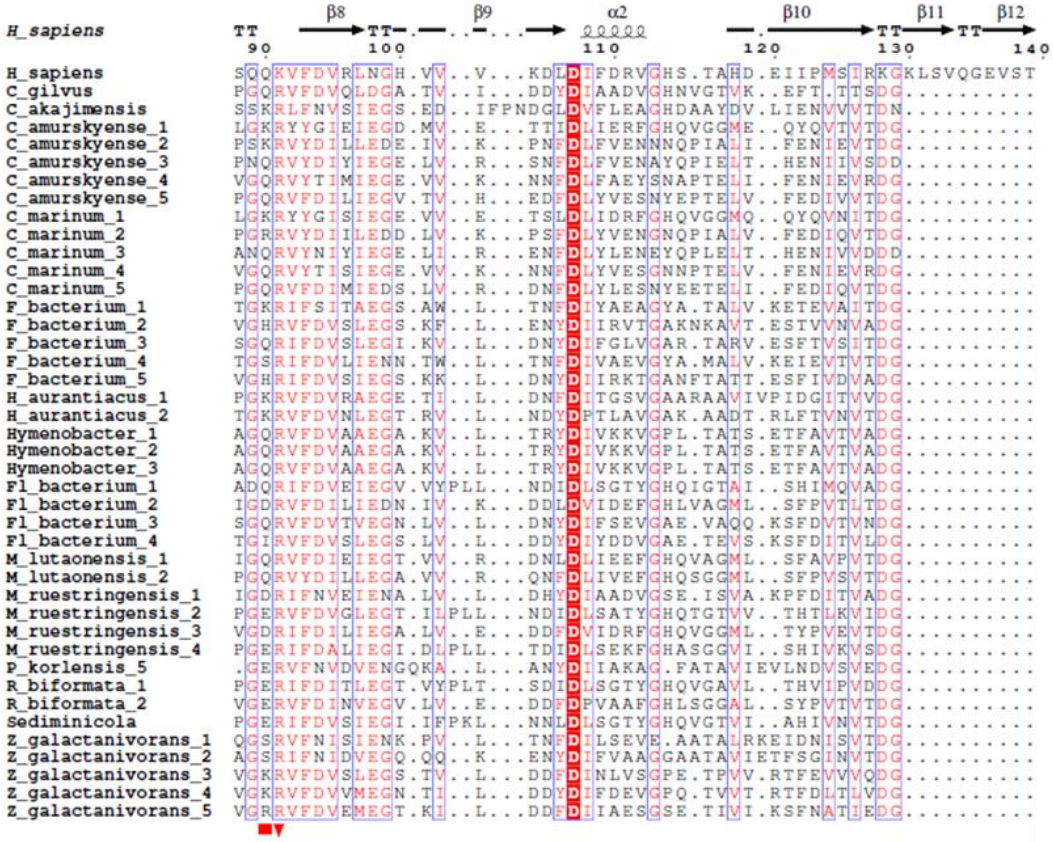

α1
β5
β6
β7
H_sapiens
40
50
60
70
80
H_sapiens
B_cavernae
C_amurskyense_1
C_amurskyense_2
C_amurskyense_3
C_amurskyense_4
C_amurskyense_5
C_marinum_1
C_marinum_2
C_marinum_3
C_marinum_4
C_marinum_5
C_marinum_6
F_bacterium_1
F_bacterium_2
F_bacterium_3
H_aurantiacus_1
H_aurantiacus_2
H_aurantiacus_3
Fl_bacterium_1
Fl_bacterium_2
Fl_bacterium_3
Fl_bacterium_4
M_dichloromethanicum_1
M_dichloromethanicum_2
M_dichloromethanicum_3
M_dichloromethanicum_4
M_extorquens_1
M_extorquens_2
M_extorquens_3
M_extorquens_4
M_chloromethanicum_1
M_chloromethanicum_2
M_chloromethanicum_3
M_chloromethanicum_4
M_populi_1
M_populi_2
M_populi_3
M_spAMS5_1
M_spAMS5_2
M_spAMS5_3
M_lutaonensis_1
M_lutaonensis_2
M_rubrisediminis_1
M_rubrisediminis_2
M_rubrisediminis_3
M_rubrisediminis_4
P_korlensis
Pseudanabaena_1
Pseudanabaena_2
Pseudanabaena_3
Pseudanabaena_4
Pseudanabaena_5
R_marinus
R_biformata_1
R_biformata_2
Sediminicola
Z_galactanivorans_1
Z_galactanivorans_2
Z_galactanivorans_3
Z_galactanivorans_4
Z_galactanivorans_5

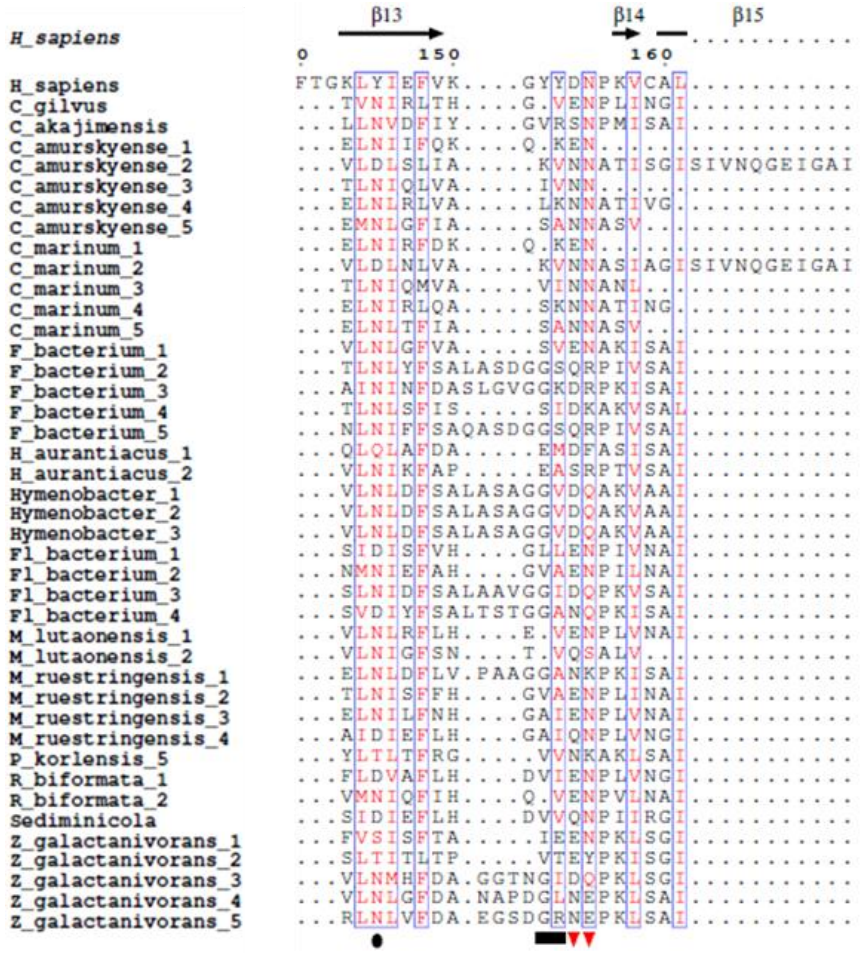

H_sapiens
β8
β9
α2
β10
90
100
110
120
H_sapiens
B_cavernae
C_amurskyense_1
C_amurskyense_2
C_amurskyense_3
C_amurskyense_4
C_amurskyense_5
C_marinum_1
C_marinum_2
C_marinum_3
C_marinum_4
C_marinum_5
C_marinum_6
F_bacterium_1
F_bacterium_2
F_bacterium_3
H_aurantiacus_1
H_aurantiacus_2
H_aurantiacus_3
Fl_bacterium_1
Fl_bacterium_2
Fl_bacterium_3
Fl_bacterium_4
M_dichloromethanicum_1
M_dichloromethanicum_2
M_dichloromethanicum_3
M_dichloromethanicum_4
M_extorquens_1
M_extorquens_2
M_extorquens_3
M_extorquens_4
M_chloromethanicum_1
M_chloromethanicum_2
M_chloromethanicum_3
M_chloromethanicum_4
M_populi_1
M_populi_2
M_populi_3
M_spAMS5_1
M_spAMS5_2
M_spAMS5_3
M_lutaonensis_1
M_lutaonensis_2
M_ruestringensis_1
M_ruestringensis_2
M_ruestringensis_3
M_ruestringensis_4
P_korlensis
Pseudanabaena_1
Pseudanabaena_2
Pseudanabaena_3
Pseudanabaena_4
Pseudanabaena_5
R_marinus
R_biformata_1
R_biformata_2
Sediminicola
Z_galactanivorans_1
Z_galactanivorans_2
Z_galactanivorans_3
Z_galactanivorans_4
Z_galactanivorans_5



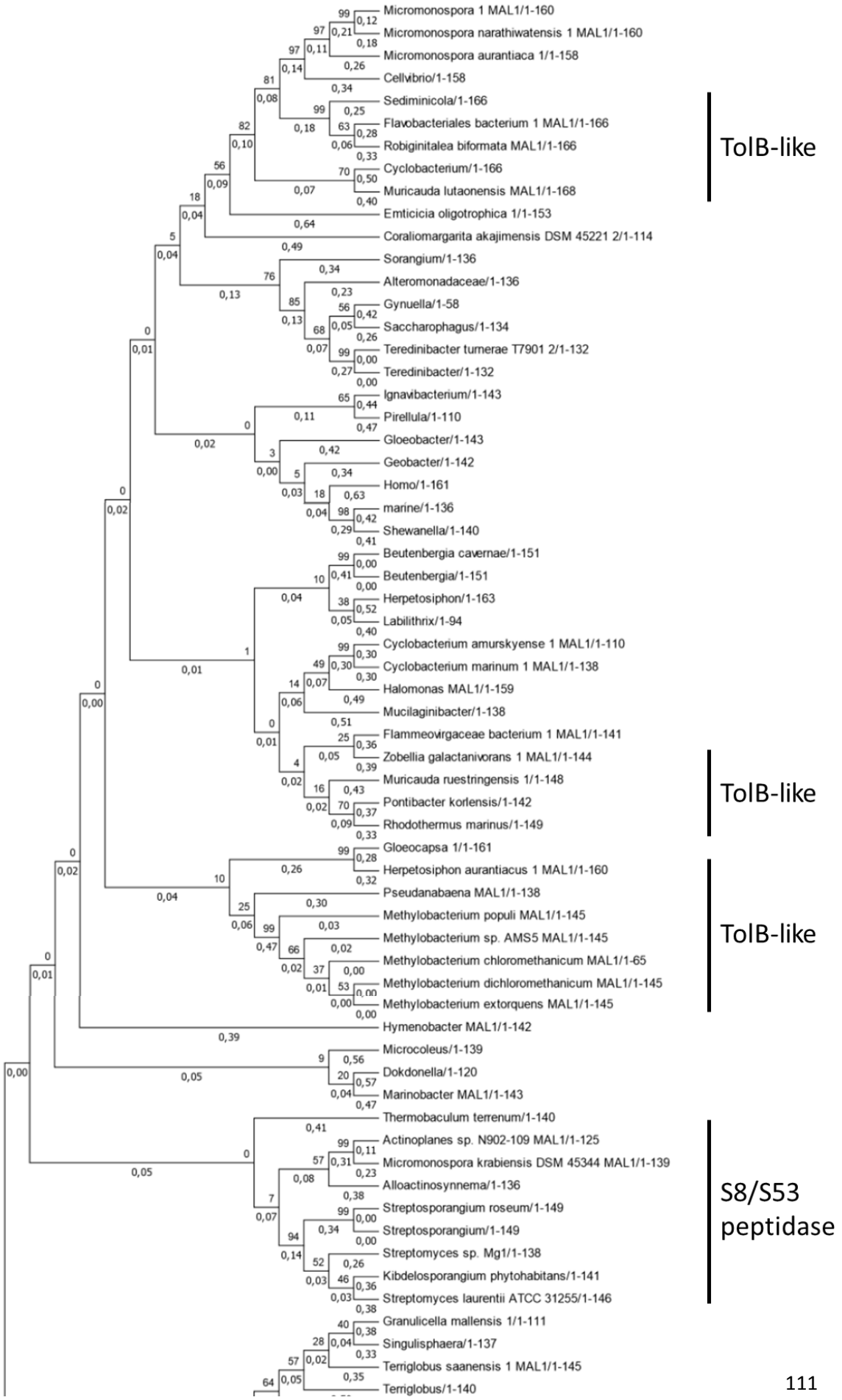
Index Figure 4- Alignment of CBM57 modules associated with TolB-like compared to the human malectin, using ClustalOmega [Sievers, *et al*, 2011] and ESPript [Robert, *et al*, 2014] programs. The amino acids in red are conserved. The red symbols above the alignment represents the binding sites residues; Red triangles-by direct hydrogen bonds; Red squares-by hydrogen bonds mediated by water; red stars- by π/CH interactions. The black symbols mark the carbohydrate-interacting residues from malectin putative binding site.

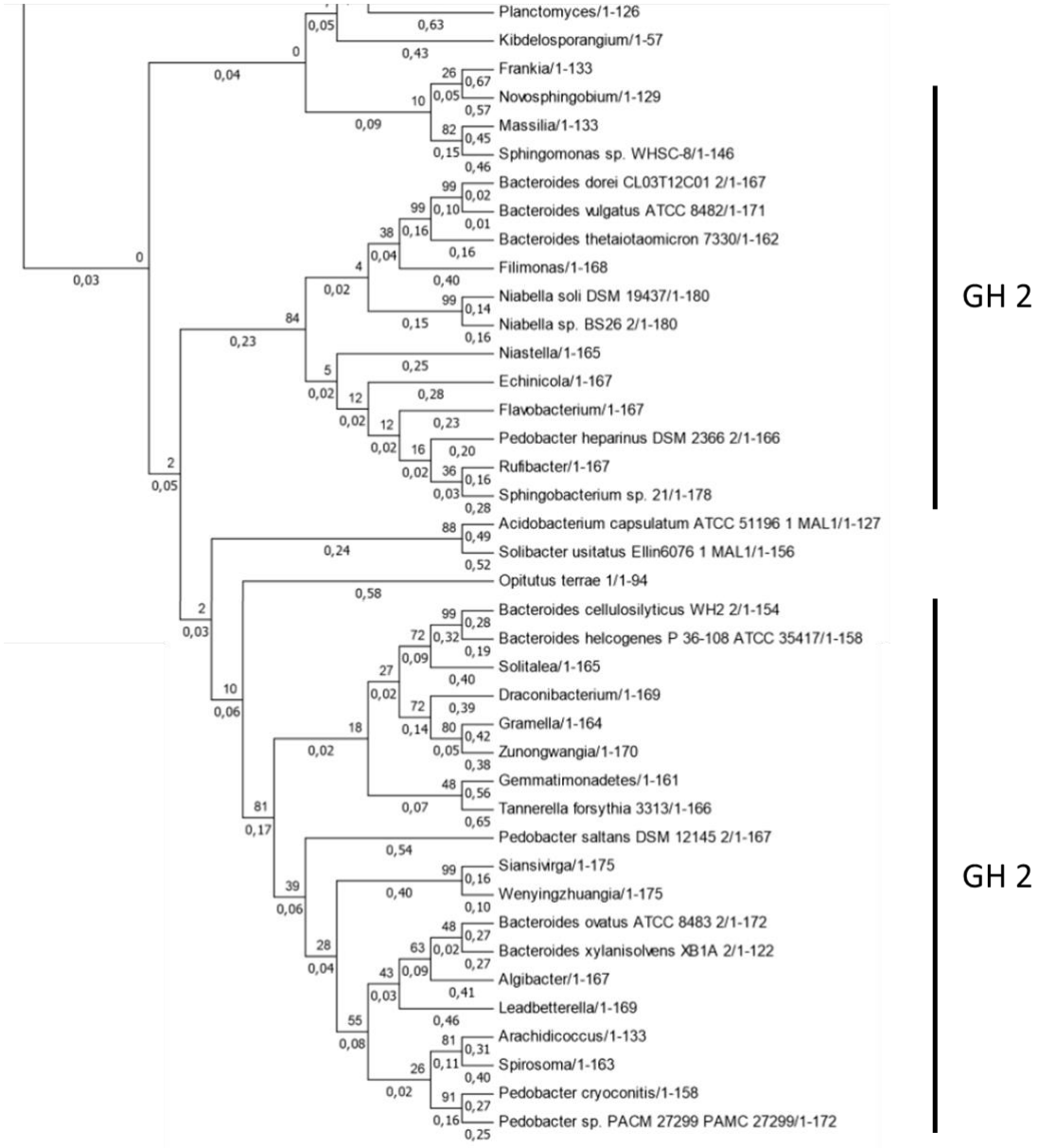

H_sapiens
H_sapiens
C_gilvus
C_akajimensis
C_amurskyense_1
C_amurskyense_2
C_amurskyense_3
C_amurskyense_4
C_amurskyense_5
C_marinum_1
C_marinum_2
C_marinum_3
C_marinum_4
C_marinum_5
F_bacterium_1
F_bacterium_2
F_bacterium_3
F_bacterium_4
F_bacterium_5
H_aurantiacus_1
H_aurantiacus_2
Hymenobacter_1
Hymenobacter_2
Hymenobacter_3
Fl_bacterium_1
Fl_bacterium_2
Fl_bacterium_3
Fl_bacterium_4
M_lutaonensis_1
M_lutaonensis_2
M_ruestringensis_1
M_ruestringensis_2
M_ruestringensis_3
M_ruestringensis_4
P_korlensis_5
R_biformata_1
R_biformata_2
Sediminicola
Z_galactanivorans_1
Z_galactanivorans_2
Z_galactanivorans_3
Z_galactanivorans_4
Z_galactanivorans_5


β8
β9
α2
β10
β11
β12
H_sapiens
90
100
110
120
130
140
H_sapiens
C_gilvus
C_akajimensis
C_amurskyense_1
C_amurskyense_2
C_amurskyense_3
C_amurskyense_4
C_amurskyense_5
C_marinum_1
C_marinum_2
C_marinum_3
C_marinum_4
C_marinum_5
F_bacterium_1
F_bacterium_2
F_bacterium_3
F_bacterium_4
F_bacterium_5
H_aurantiacus_1
H_aurantiacus_2
Hymenobacter_1
Hymenobacter_2
Hymenobacter_3
Fl_bacterium_1
Fl_bacterium_2
Fl_bacterium_3
Fl_bacterium_4
M_lutaonensis_1
M_lutaonensis_2
M_ruestringensis_1
M_ruestringensis_2
M_ruestringensis_3
M_ruestringensis_4
P_korlensis_5
R_biformata_1
R_biformata_2
Sediminicola
Z_galactanivorans_1
Z_galactanivorans_2
Z_galactanivorans_3
Z_galactanivorans_4
Z_galactanivorans_5

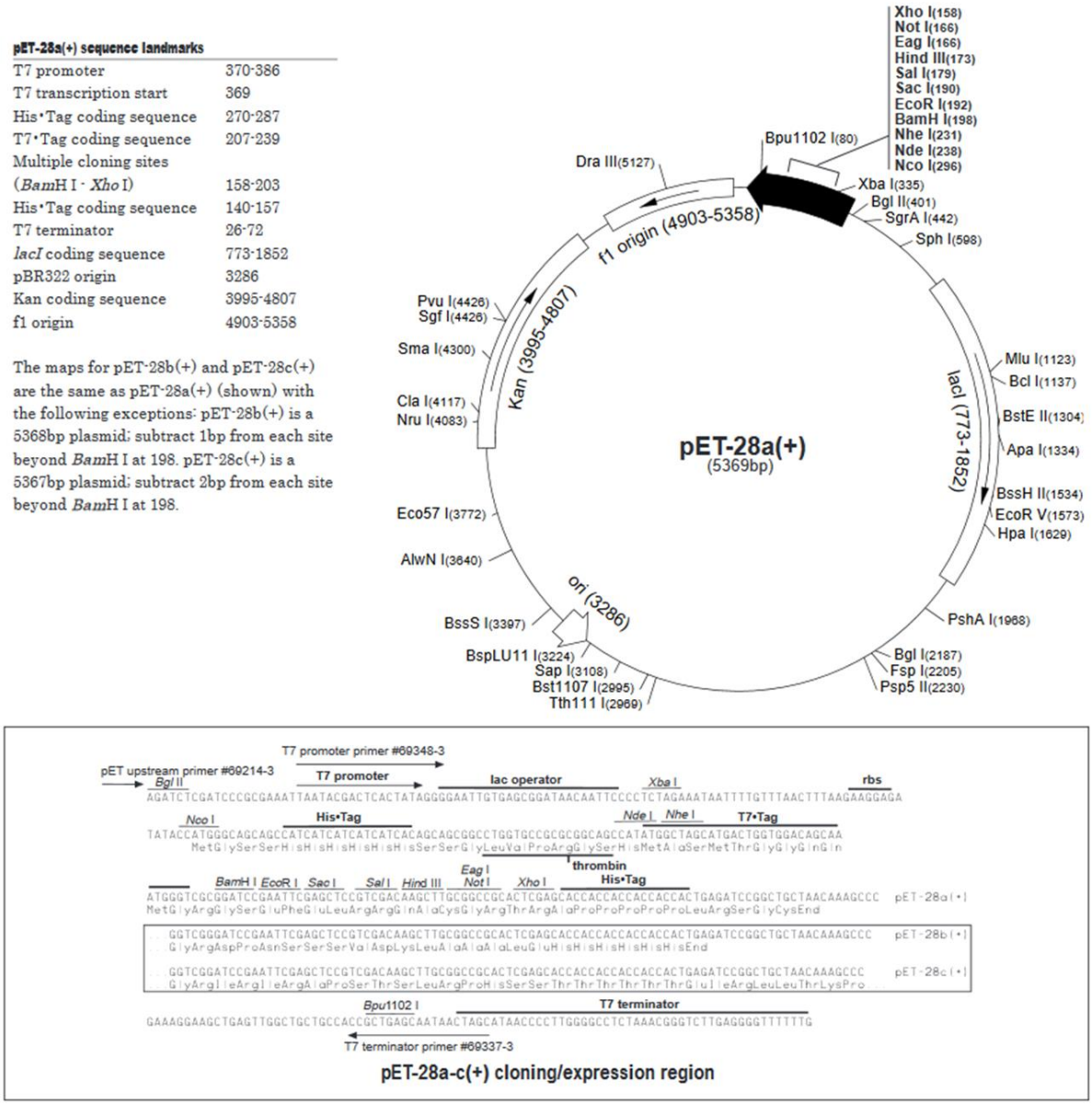


Index Figure 5- Alignment of CBM57 modules associated with PKD compared to the human malectin, using ClustalOmega [Sievers, *et al*, 2011] and ESPript [Robert, *et al*, 2014] programs. The amino acids in red are conserved. The red symbols above the alignment represents the binding sites residues; Red triangles-by direct hydrogen bonds; Red squares-by hydrogen bonds mediated by water; red stars- by π/CH interactions. The black symbols mark the carbohydrate-interacting residues from malectin putative binding site.

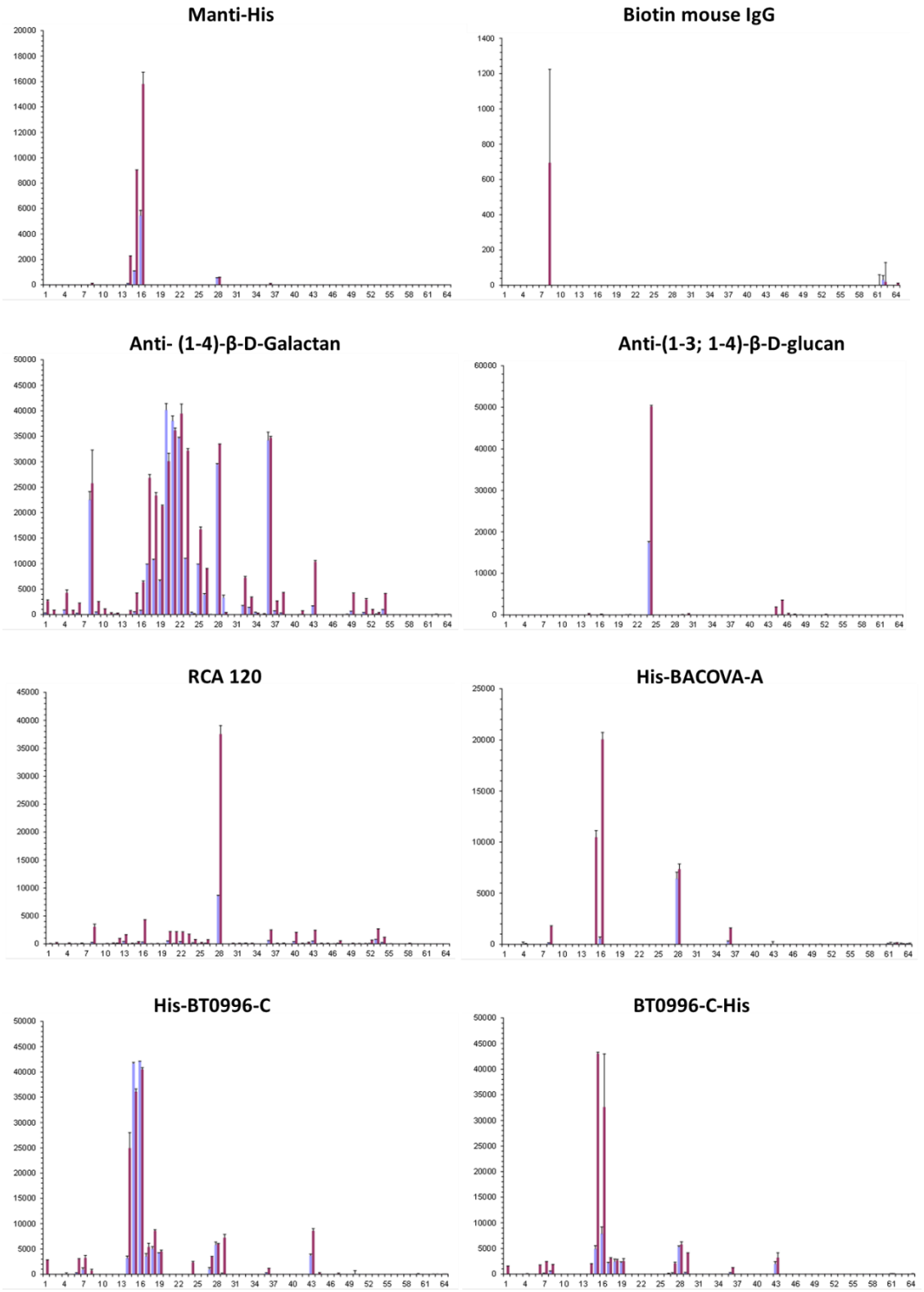

Micromonospora 1 MAL1/1-160
Micromonospora narathiwatensis 1 MAL1/1-160
Micromonospora aurantiaca 1/1-158
Cellvibrio/1-158
Sediminicola/1-166
Flavobacteriales bacterium 1 MAL1/1-166
Robiginitalea biformata MAL1/1-166
Cyclobacterium/1-166
Muricauda lutaonensis MAL1/1-168
Emticicia oligotrophica 1/1-153
Coraliomargarita akajimensis DSM 45221 2/1-114
Sorangium/1-136
Alteromonadaceae/1-136
Gynuella/1-58
Saccharophagus/1-134
Teredinibacter turnerae T7901 2/1-132
Teredinibacter/1-132
Ignavibacterium/1-143
Pirellula/1-110
Gloeobacter/1-143
Geobacter/1-142
Homo/1-161
marine/1-136
Shewanella/1-140
Beutenbergia cavernae/1-151
Beutenbergia/1-151
Herpetosiphon/1-163
Labilithrix/1-94
Cyclobacterium amurskyense 1 MAL1/1-110
Cyclobacterium marinum 1 MAL1/1-138
Halomonas MAL1/1-159
Mucilaginibacter/1-138
Flammeovirgaceae bacterium 1 MAL1/1-141
Zobellia galactanivorans 1 MAL1/1-144
Muricauda ruestringensis 1/1-148
Pontibacter korlensis/1-142
Rhodothermus marinus/1-149
Gloeocapsa 1/1-161
Herpetosiphon aurantiacus 1 MAL1/1-160
Pseudanabaena MAL1/1-138
Methylobacterium populi MAL1/1-145
Methylobacterium sp. AMS5 MAL1/1-145
Methylobacterium chloromethanicum MAL1/1-65
Methylobacterium dichloromethanicum MAL1/1-145
Methylobacterium extorquens MAL1/1-145
Hymenobacter MAL1/1-142
Microcoleus/1-139
Dokdonella/1-120
Marinobacter MAL1/1-143
Thermobaculum terrenum/1-140
Actinoplanes sp. N902-109 MAL1/1-125
Micromonospora krabiensis DSM 45344 MAL1/1-139
Alloactinosynnema/1-136
Streptosporangium roseum/1-149
Streptosporangium/1-149
Streptomyces sp. Mg1/1-138
Kibdelosporangium phytohabitans/1-141
Streptomyces laurentii ATCC 31255/1-146
Granulicella mallensis 1/1-111
Singulisphaera/1-137
Terriglobus saanensis 1 MAL1/1-145
Terriglobus/1-140
TolB-like
TolB-like
TolB-like
S8/S53 peptidase



Index Figure 6- Phylogenetic tree construction for one malectin-like module in each bacteria specie using MEGA7 program. The method used was Neighbor-joining. Gaps presented in amino acids sequences were treated using the pairwise-deletion option. The validation of phylogenetic tree was done by performing 1000 bootstrap replications.



Index Figure 7- The pET28 map, that carry an N-terminal His-tag and an C-terminal His-tag, the T7 promotor, the selection marker for kanamycin and the respective restriction sites. Figure taken from https://www.staff.ncl.ac.uk/p.dean/pET28.pdf



Index Figure 8- Raw glycan microarray data analysis of proteins for que quality control of the microarray set and of proteins for characterization studies. Protein names are written on top of each graph and each recognize different epitopes and have different specificities (index table 3). Glycans sequence information of the probes included in the microarray are in Index Table 2. The interacting signals are the fluorescence intensities of duplicate spots of probe with error bars. Each probe was printed at two concentrations: 0.1 and 0.5 mg (dry weight). Quatified fluorencense intensity is plotted on the y-axis. Glycan probes are plotted on the x-axis.

A

Tube #	Salt	Tube #	Polymer	Tube #	pH ◊
1.	0.2 M Sodium fluoride	1.	20% w/v Polyethylene glycol 3,350	1.	7.3
2.	0.2 M Potassium fluoride	2.	20% w/v Polyethylene glycol 3,350	2.	7.3
3.	0.2 M Ammonium fluoride	3.	20% w/v Polyethylene glycol 3,350	3.	6.2
4.	0.2 M Lithium chloride	4.	20% w/v Polyethylene glycol 3,350	4.	6.8
5.	0.2 M Magnesium chloride hexahydrate	5.	20% w/v Polyethylene glycol 3,350	5.	5.9
6.	0.2 M Sodium chloride	6.	20% w/v Polyethylene glycol 3,350	6.	6.9
7.	0.2 M Calcium chloride dihydrate	7.	20% w/v Polyethylene glycol 3,350	7.	5.1
8.	0.2 M Potassium chloride	8.	20% w/v Polyethylene glycol 3,350	8.	7.0
9.	0.2 M Ammonium chloride	9.	20% w/v Polyethylene glycol 3,350	9.	6.3
10.	0.2 M Sodium iodide	10.	20% w/v Polyethylene glycol 3,350	10.	7.0
11.	0.2 M Potassium iodide	11.	20% w/v Polyethylene glycol 3,350	11.	7.0
12.	0.2 M Ammonium iodide	12.	20% w/v Polyethylene glycol 3,350	12.	6.2
13.	0.2 M Sodium thiocyanate	13.	20% w/v Polyethylene glycol 3,350	13.	6.9
14.	0.2 M Potassium thiocyanate	14.	20% w/v Polyethylene glycol 3,350	14.	7.0
15.	0.2 M Lithium nitrate	15.	20% w/v Polyethylene glycol 3,350	15.	7.1
16.	0.2 M Magnesium nitrate hexahydrate	16.	20% w/v Polyethylene glycol 3,350	16.	5.9
17.	0.2 M Sodium nitrate	17.	20% w/v Polyethylene glycol 3,350	17.	6.8
18.	0.2 M Potassium nitrate	18.	20% w/v Polyethylene glycol 3,350	18.	6.8
19.	0.2 M Ammonium nitrate	19.	20% w/v Polyethylene glycol 3,350	19.	6.2
20.	0.2 M Magnesium formate dihydrate	20.	20% w/v Polyethylene glycol 3,350	20.	7.0
21.	0.2 M Sodium formate	21.	20% w/v Polyethylene glycol 3,350	21.	7.2
22.	0.2 M Potassium formate	22.	20% w/v Polyethylene glycol 3,350	22.	7.3
23.	0.2 M Ammonium formate	23.	20% w/v Polyethylene glycol 3,350	23.	6.6
24.	0.2 M Lithium acetate dihydrate	24.	20% w/v Polyethylene glycol 3,350	24.	7.9
25.	0.2 M Magnesium acetate tetrahydrate	25.	20% w/v Polyethylene glycol 3,350	25.	7.9
26.	0.2 M Zinc acetate dihydrate	26.	20% w/v Polyethylene glycol 3,350	26.	6.4
27.	0.2 M Sodium acetate trihydrate	27.	20% w/v Polyethylene glycol 3,350	27.	8.0
28.	0.2 M Calcium acetate hydrate	28.	20% w/v Polyethylene glycol 3,350	28.	7.5
29.	0.2 M Potassium acetate	29.	20% w/v Polyethylene glycol 3,350	29.	8.1
30.	0.2 M Ammonium acetate	30.	20% w/v Polyethylene glycol 3,350	30.	7.1
31.	0.2 M Lithium sulfate monohydrate	31.	20% w/v Polyethylene glycol 3,350	31.	6.0
32.	0.2 M Magnesium sulfate heptahydrate	32.	20% w/v Polyethylene glycol 3,350	32.	6.0
33.	0.2 M Sodium sulfate decahydrate	33.	20% w/v Polyethylene glycol 3,350	33.	6.7
34.	0.2 M Potassium sulfate	34.	20% w/v Polyethylene glycol 3,350	34.	6.8
35.	0.2 M Ammonium sulfate	35.	20% w/v Polyethylene glycol 3,350	35.	6.0
36.	0.2 M Sodium tartrate dibasic dihydrate	36.	20% w/v Polyethylene glycol 3,350	36.	7.3
37.	0.2 M Potassium sodium tartrate tetrahydrate	37.	20% w/v Polyethylene glycol 3,350	37.	7.4
38.	0.2 M Ammonium tartrate dibasic	38.	20% w/v Polyethylene glycol 3,350	38.	6.6
39.	0.2 M Sodium phosphate monobasic monohydrate	39.	20% w/v Polyethylene glycol 3,350	39.	4.7
40.	0.2 M Sodium phosphate dibasic dihydrate	40.	20% w/v Polyethylene glycol 3,350	40.	9.1
41.	0.2 M Potassium phosphate monobasic	41.	20% w/v Polyethylene glycol 3,350	41.	4.8
42.	0.2 M Potassium phosphate dibasic	42.	20% w/v Polyethylene glycol 3,350	42.	9.2
43.	0.2 M Ammonium phosphate monobasic	43.	20% w/v Polyethylene glycol 3,350	43.	4.6
44.	0.2 M Ammonium phosphate dibasic	44.	20% w/v Polyethylene glycol 3,350	44.	8.0
45.	0.2 M Lithium citrate tribasic tetrahydrate	45.	20% w/v Polyethylene glycol 3,350	45.	8.4
46.	0.2 M Sodium citrate tribasic dihydrate	46.	20% w/v Polyethylene glycol 3,350	46.	8.3
47.	0.2 M Potassium citrate tribasic monohydrate	47.	20% w/v Polyethylene glycol 3,350	47.	8.3
48.	0.2 M Ammonium citrate dibasic	48.	20% w/v Polyethylene glycol 3,350	48.	5.1

B

Tube #	Salt	Buffer ◊	Polymer
1.	0.1 M Sodium malonate pH 4.0	None	12% w/v Polyethylene glycol 3,350
2.	0.2 M Sodium malonate pH 4.0	None	20% w/v Polyethylene glycol 3,350
3.	0.1 M Sodium malonate pH 5.0	None	12% w/v Polyethylene glycol 3,350
4.	0.2 M Sodium malonate pH 5.0	None	20% w/v Polyethylene glycol 3,350
5.	0.1 M Sodium malonate pH 6.0	None	12% w/v Polyethylene glycol 3,350
6.	0.2 M Sodium malonate pH 6.0	None	20% w/v Polyethylene glycol 3,350
7.	0.1 M Sodium malonate pH 7.0	None	12% w/v Polyethylene glycol 3,350
8.	0.2 M Sodium malonate pH 7.0	None	20% w/v Polyethylene glycol 3,350
9.	4% v/v Tacsimate™ pH 4.0	None	12% w/v Polyethylene glycol 3,350
10.	8% v/v Tacsimate™ pH 4.0	None	20% w/v Polyethylene glycol 3,350
11.	4% v/v Tacsimate™ pH 5.0	None	12% w/v Polyethylene glycol 3,350
12.	8% v/v Tacsimate™ pH 5.0	None	20% w/v Polyethylene glycol 3,350
13.	4% v/v Tacsimate™ pH 6.0	None	12% w/v Polyethylene glycol 3,350
14.	8% v/v Tacsimate™ pH 6.0	None	20% w/v Polyethylene glycol 3,350
15.	4% v/v Tacsimate™ pH 7.0	None	12% w/v Polyethylene glycol 3,350
16.	8% v/v Tacsimate™ pH 7.0	None	20% w/v Polyethylene glycol 3,350
17.	4% v/v Tacsimate™ pH 8.0	None	12% w/v Polyethylene glycol 3,350
18.	8% v/v Tacsimate™ pH 8.0	None	20% w/v Polyethylene glycol 3,350
19.	0.1 M Succinic acid pH 7.0	None	12% w/v Polyethylene glycol 3,350
20.	0.2 M Succinic acid pH 7.0	None	20% w/v Polyethylene glycol 3,350
21.	0.1 M Ammonium citrate tribasic pH 7.0	None	12% w/v Polyethylene glycol 3,350
22.	0.2 M Ammonium citrate tribasic pH 7.0	None	20% w/v Polyethylene glycol 3,350
23.	0.1 M DL-Malic acid pH 7.0	None	12% w/v Polyethylene glycol 3,350
24.	0.2 M DL-Malic acid pH 7.0	None	20% w/v Polyethylene glycol 3,350
25.	0.1 M Sodium acetate trihydrate pH 7.0	None	12% w/v Polyethylene glycol 3,350
26.	0.2 M Sodium acetate trihydrate pH 7.0	None	20% w/v Polyethylene glycol 3,350
27.	0.1 M Sodium formate pH 7.0	None	12% w/v Polyethylene glycol 3,350
28.	0.2 M Sodium formate pH 7.0	None	20% w/v Polyethylene glycol 3,350
29.	0.1 M Ammonium tartrate dibasic pH 7.0	None	12% w/v Polyethylene glycol 3,350
30.	0.2 M Ammonium tartrate dibasic pH 7.0	None	20% w/v Polyethylene glycol 3,350
31.	2% v/v Tacsimate™ pH 4.0	0.1 M Sodium acetate trihydrate pH 4.6	16% w/v Polyethylene glycol 3,350
32.	2% v/v Tacsimate™ pH 5.0	0.1 M Sodium citrate tribasic dihydrate pH 5.6	16% w/v Polyethylene glycol 3,350
33.	2% v/v Tacsimate™ pH 6.0	0.1 M BIS-TRIS pH 6.5	20% w/v Polyethylene glycol 3,350
34.	2% v/v Tacsimate™ pH 7.0	0.1 M HEPES pH 7.5	20% w/v Polyethylene glycol 3,350
35.	2% v/v Tacsimate™ pH 8.0	0.1 M Tris pH 8.5	16% w/v Polyethylene glycol 3,350
36.	None	0.07 M Citric acid, 0.03 M BIS-TRIS propane / pH 3.4	16% w/v Polyethylene glycol 3,350
37.	None	0.06 M Citric acid, 0.04 M BIS-TRIS propane / pH 4.1	16% w/v Polyethylene glycol 3,350
38.	None	0.05 M Citric acid, 0.05 M BIS-TRIS propane / pH 5.0	16% w/v Polyethylene glycol 3,350
39.	None	0.04 M Citric acid, 0.06 M BIS-TRIS propane / pH 6.4	20% w/v Polyethylene glycol 3,350
40.	None	0.03 M Citric acid, 0.07 M BIS-TRIS propane / pH 7.6	20% w/v Polyethylene glycol 3,350
41.	None	0.02 M Citric acid, 0.08 M BIS-TRIS propane / pH 8.8	16% w/v Polyethylene glycol 3,350
42.	0.02 M Calcium chloride dihydrate, 0.02 M Cadmium chloride hydrate, 0.02 M Cobalt(II) chloride hexahydrate	None	20% w/v Polyethylene glycol 3,350
43.	0.01 M Magnesium chloride hexahydrate 0.005 M Nickel(II) chloride hexahydrate	0.1 M HEPES sodium pH 7.0	15% w/v Polyethylene glycol 3,350
44.	0.02 M Zinc chloride	None	20% w/v Polyethylene glycol 3,350
45.	0.15 M Cesium chloride	None	15% w/v Polyethylene glycol 3,350
46.	0.2 M Sodium bromide	None	20% w/v Polyethylene glycol 3,350
47.	1% w/v Tryptone, 0.001 M Sodium azide	0.05 M HEPES sodium pH 7.0	12% w/v Polyethylene glycol 3,350
48.	1% w/v Tryptone, 0.001 M Sodium azide	0.05 M HEPES sodium pH 7.0	20% w/v Polyethylene glycol 3,350

Index Figure 9- PEG Ion (A) and PEG Ion2 (B) commercial screens from Hampton Research used in crystallization trials. Figure taken from www.hamptonresearch.com.

A

Tube #	Conc.	Units	Salt 1	Conc.	Units	Buffer	pH	Conc.	Units	Precipitant 1
1	0.02	M	Calcium chloride dihydrate	0.1	M	Sodium acetate	4.6	30	% v/v	MPD
2	0.2	M	Ammonium acetate	0.1	M	Sodium acetate	4.6	30	% w/v	PEG 4000
3	0.2	M	Ammonium sulfate	0.1	M	Sodium acetate	4.6	25	% w/v	PEG 4000
4	2.0	M	Sodium formate	0.1	M	Sodium acetate	4.6			
5	2.0	M	Ammonium sulfate	0.1	M	Sodium acetate	4.6			
6				0.1	M	Sodium acetate	4.6	8	% w/v	PEG 4000
7	0.2	M	Ammonium acetate	0.1	M	Sodium citrate	5.6	30	% w/v	PEG 4000
8	0.2	M	Ammonium acetate	0.1	M	Sodium citrate	5.6	30	% v/v	MPD
9				0.1	M	Sodium citrate	5.6	20	% w/v	PEG 4000/
								20	% v/v	2-Propanol
10	1.0	M	Ammonium phosphate monobasic	0.1	M	Sodium citrate	5.6			
11	0.2	M	Calcium chloride dihydrate	0.1	M	Sodium acetate	4.6	20	% v/v	2-Propanol
12	1.4	M	Sodium acetate trihydrate	0.1	M	Sodium cacodylate	6.5			
13	0.2	M	Sodium citrate tribasic dihydrate	0.1	M	Sodium cacodylate	6.5	30	% v/v	2-Propanol
14	0.2	M	Ammonium sulfate	0.1	M	Sodium cacodylate	6.5	30	% w/v	PEG 8000
15	0.2	M	Magnesium acetate tetrahydrate	0.1	M	Sodium cacodylate	6.5	20	% w/v	PEG 8000
16	0.2	M	Magnesium acetate tetrahydrate	0.1	M	Sodium cacodylate	6.5	30	% v/v	MPD
17	1.0	M	Sodium acetate trihydrate	0.1	M	Imidazole	6.5			
18	0.2	M	Sodium acetate trihydrate	0.1	M	Sodium cacodylate	6.5	30	% w/v	PEG 8000
19	0.2	M	Zinc acetate dihydrate	0.1	M	Sodium cacodylate	6.5	18	% w/v	PEG 8000
20	0.2	M	Calcium acetate hydrate	0.1	M	Sodium cacodylate	6.5	18	% w/v	PEG 8000
21	0.2	M	Sodium citrate tribasic dihydrate	0.1	M	Sodium HEPES	7.5	30	% v/v	MPD
22	0.2	M	Magnesium chloride hexahydrate	0.1	M	Sodium HEPES	7.5	30	% v/v	2-Propanol
23	0.2	M	Calcium chloride dihydrate	0.1	M	Sodium HEPES	7.5	28	% v/v	PEG 400
24	0.2	M	Magnesium chloride hexahydrate	0.1	M	Sodium HEPES	7.5	30	% v/v	PEG 400
25	0.2	M	Sodium citrate tribasic dihydrate	0.1	M	Sodium HEPES	7.5	20	% v/v	2-Propanol
26	0.8	M	Potasium sodium tartrate tetrahydrate	0.1	M	Sodium HEPES	7.5			
27	1.5	M	Lithium sulfate	0.1	M	Sodium HEPES	7.5			
28	0.8	M	Sodium phosphate monobasic monohydrate/	0.1	M	Sodium HEPES	7.5			
	0.8	M	Potasium phosphate monobasic							
29	1.4	M	Sodium citrate tribasic dihydrate	0.1	M	Sodium HEPES	7.5			
30	2.0	M	Ammonium sulfate	0.1	M	Sodium HEPES	7.5	2	% v/v	PEG 400
31				0.1	M	Sodium HEPES	7.5	20	% w/v	PEG 4000/
								10	% v/v	2-Propanol
32	2.0	M	Ammonium sulfate	0.1	M	Tris	8.5			
33	0.2	M	Magnesium chloride hexahydrate	0.1	M	Tris	8.5	30	% w/v	PEG 4000
34	0.2	M	Sodium citrate tribasic dihydrate	0.1	M	Tris	8.5	30	% v/v	PEG 400
35	0.2	M	Lithium sulfate	0.1	M	Tris	8.5	30	% w/v	PEG 4000
36	0.2	M	Ammonium acetate	0.1	M	Tris	8.5	30	% v/v	2-Propanol
37	0.2	M	Sodium acetate trihydrate	0.1	M	Tris	8.5	30	% w/v	PEG 4000
38				0.1	M	Tris	8.5	8	% w/v	PEG 8000
39	2.0	M	Ammonium phosphate monobasic	0.1	M	Tris	8.5			
40	0.4	M	Potasium sodium tartrate tetrahydrate							
41	0.4	M	Ammonium phosphate monobasic							
42	0.2	M	Ammonium sulfate					30	% w/v	PEG 8000
43	0.2	M	Ammonium sulfate					30	% w/v	PEG 4000
44	2.0	M	Ammonium sulfate							
45	4.0	M	Sodium formate							
46	0.05	M	Potassium phosphate monobasic							
47								30	% w/v	PEG 1500
48	0.2	M	Magnesium formate dihydrate							
49	1.0	M	Lithium sulfate					2	% w/v	PEG 8000
50	0.5	M	Lithium sulfate					15	% w/v	PEG 8000

B

Tube #	Conc.	Salt	Conc.	Buffer	pH	Conc.	Precipitant
1	0.1 M	Sodium chloride	0.1 M	BICINE	9.0	30 % v/v	PEG 500 MME
2	2.0 M	Magnesium chloride hexahydrate	0.1 M	BICINE	9.0		
3			0.1 M	BICINE	9.0	10 % w/v	PEG 20000
						2 % v/v	1,4-Dioxane
4	0.2 M	Magnesium chloride hexahydrate	0.1 M	Tris	8.5	3.4 M	1,6-Hexanediol
5			0.1 M	Tris	8.5	25 % v/v	tert-Butanol
6	1.0 M	Lithium sulfate	0.1 M	Tris	8.5		
	0.01 M	Nickel(II) chloride hexahydrate					
7	1.5 M	Ammonium sulfate	0.1 M	Tris	8.5	12 % v/v	Glycerol
8	0.2 M	Ammonium phosphate monobasic	0.1 M	Tris	8.5	50 % v/v	MPD
9			0.1 M	Tris	8.5	20 % v/v	Ethanol
10	0.01 M	Nickel(II) chloride hexahydrate	0.1 M	Tris	8.5	20 % w/v	PEG 2000 MME
11	0.5 M	Ammonium sulfate	0.1 M	Sodium HEPES	7.5	30 % v/v	MPD
12			0.1 M	Sodium HEPES	7.5	10 % w/v	PEG 6000
						5 % v/v	MPD
13			0.1 M	Sodium HEPES	7.5	20 % v/v	Jeffamine® M-600
14	1.6 M	Ammonium sulfate	0.1 M	Sodium HEPES	7.5		
	0.1 M	Sodium chloride					
15	2.0 M	Ammonium formate	0.1 M	Sodium HEPES	7.5		
16	1.0 M	Sodium acetate trihydrate	0.1 M	Sodium HEPES	7.5		
	0.05 M	Cadmium sulfate $^8/_3$-hydrate					
17			0.1 M	Sodium HEPES	7.5	70 % v/v	MPD
18	4.3 M	Sodium chloride	0.1 M	Sodium HEPES	7.5		
19			0.1 M	Sodium HEPES	7.5	10 % w/v	PEG 8000
						8 % v/v	Ethylene glycol
20	1.6 M	Magnesium sulfate heptahydrate	0.1 M	MES	6.5		
21	2.0 M	Sodium chloride	0.1 M	MES	6.5		
	0.1 M	Potassium phosphate monobasic					
	0.1 M	Sodium phosphate monobasic monohydrate					
22			0.1 M	MES	6.5	12 % w/v	PEG 20000
23	1.6 M	Ammonium sulfate	0.1 M	MES	6.5	10 % v/v	1,4-Dioxane
24	0.05 M	Cesium chloride	0.1 M	MES	6.5	30 % v/v	Jeffamine® M-600
25	0.01 M	Cobalt(II) chloride hexahydrate	0.1 M	MES	6.5		
	1.8 M	Ammonium sulfate					
26	0.2 M	Ammonium sulfate	0.1 M	MES	6.5	30 % w/v	PEG 5000 MME
27	0.01 M	Zinc sulfate heptahydrate	0.1 M	MES	6.5	25 % v/v	PEG 500 MME
28			0.1 M	Sodium HEPES	7.5	20 % w/v	PEG 10000
29	2.0 M	Ammonium sulfate	0.1 M	Sodium citrate	5.6		
	0.2 M	Potassium sodium tartrate tetrahydrate					
30	1.0 M	Lithium sulfate	0.1 M	Sodium citrate	5.6		
	0.5 M	Ammonium sulfate					
31	0.5 M	Sodium chloride	0.1 M	Sodium citrate	5.6	4 % v/v	Polyethyleneimine
32			0.1 M	Sodium citrate	5.6	35 % v/v	tert-Butanol
33	0.01 M	Iron(III) chloride hexahydrate	0.1 M	Sodium citrate	5.6	10 % v/v	Jeffamine® M-600
34	0.01 M	Manganese(II) chloride tetrahydrate	0.1 M	Sodium citrate	5.6	2.5 M	1,6-Hexanediol
35	2.0 M	Sodium chloride	0.1 M	Sodium acetate	4.6		
36	0.2 M	Sodium chloride	0.1 M	Sodium acetate	4.6	30 % v/v	MPD
37	0.01 M	Cobalt(II) chloride hexahydrate	0.1 M	Sodium acetate	4.6	1.0 M	1,6-Hexanediol
38	0.1 M	Cadmium chloride hemi(pentahydrate)	0.1 M	Sodium acetate	4.6	30 % v/v	PEG 400
39	0.2 M	Ammonium sulfate	0.1 M	Sodium acetate	4.6	30 % w/v	PEG 2000 MME
40	2.0 M	Sodium chloride				10 % w/v	PEG 6000
41	0.5 M	Sodium chloride					
	0.1 M	Magnesium chloride hexahydrate					
	0.01 M	CTAB					
42						25 % v/v	Ethylene glycol
43						35 % v/v	1,4-Dioxane
44	2.0 M	Ammonium sulfate				5 % v/v	2-Propanol
45			1.0 M	Imidazole	7.0		
46						10 % w/v	PEG 1000
						10 % w/v	PEG 8000
47	1.5 M	Sodium chloride				10 % v/v	Ethanol
48			1.6 M	Sodium citrate	6.5		
49						15 % w/v	Polyvinylpyrrolidone
50	2.0 M	Urea					

Index Figure 10- Structure 1 (A) and Structure 2 (B) commercial screens from Molecular Devices used in crystallization trials. Figure taken from https://www.moleculardimensions.com.

Index Information 2: Poster Communications

Assignment of biological roles for CBM57 modules in human gut bacteria

Frederico Lourenço[1], Catarina Leote[1], Viviana G. Correia[1], Joana L.A. Brás[3], Carlos M.G.A. Fontes[2,3], Maria João Romão[1], Ana Luísa Carvalho[1], Angelina S. Palma[1], Benedita A. Pinheiro[1]

[1] UCIBIO-REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
[2] CIISA/FMV-UL, Faculdade de Medicina Veterinária, Universidade de Lisboa, Av. da Universidade Técnica, 1300-477 Lisboa, Portugal
[3] NZYTech, Lda, Estrada do Paço do Lumiar, Campus do Lumiar, 1649-038 Lisboa, Portugal

Malectin and CBM57 family

Malectin (figure 1) is a highly-conserved lectin in the animal kingdom that participates in the N-glycosylation pathway in the endoplasmic reticulum. It is highly specific to the di-glucosylated oligosaccharide ($Glc_2Man_9GlcNAc_2$), an intermediate of N-glycosylation pathway, crucial for quality control of glycoproteins folding [1].

Malectin is composed of a β-sandwich core and the binding site is located in the loop region [2]. Several **carbohydrate binding modules (CBMs)** were discovered across other kingdoms of life that share amino acid sequence homology with malectin, being grouped in CBM57 family of the Carbohydrate Active Enzymes Database (CAZy).

The CBM57 family and structurally related groups (CBM6 and CBM35) are type-C CBMs and have, like malectin, the binding pocket present in the loop region. They are expected to have diverse carbohydrate binding specificities, which are here inspected.

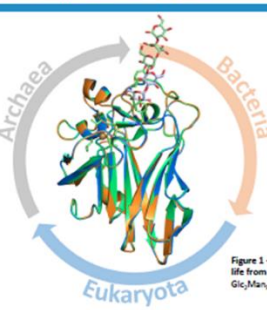


Figure 1 - **Malectin homologues are present in all kingdoms of life from Archaea to Eukaryota.** Malectin structure with $Glc_2Man_9GlcNAc_2$ solved by X-ray crystallography.

Same Specificity?

PROJECT'S AIM:

- Determine **the architecture of the proteins** from the human gut bacteria containing **CBM35** and **CBM57 modules**.
- Compare their CBM amino acid sequences.
- Determine their carbohydrate specificity using **glycan microarrays**.
- Structurally characterize the selected CBM57 and CBM35 modules using **X-ray crystallography**.

CBM57 family amino acid sequence analysis

Figure 2 - **Protein architecture predicted from both bacteria using InterproScan.** *SP* represents signal peptide. GH represents glycosidase hydrolase. UNK represents a domain unknown function. The red arrows indicate the selected CBM modules to study.

- Two proteins containing **CBM57 and CBM35 modules from the human gut bacteria *Bacteroides ovatus* and *Bacteroides thetaiotaomicron*** were retrieved from the CAZy database. Their architecture was predicted using the **InterProScan program** (Figure 2).
- Because of the presence of several similar CBM modules, each **CBM amino acid sequence was aligned to malectin using Clustal Omega and ESPript** (figure 3) in order to predict possible different carbohydrate specificities.
- Despite predicted secondary structure conservation, **malectin glycan-interacting amino acid residues are not conserved**.

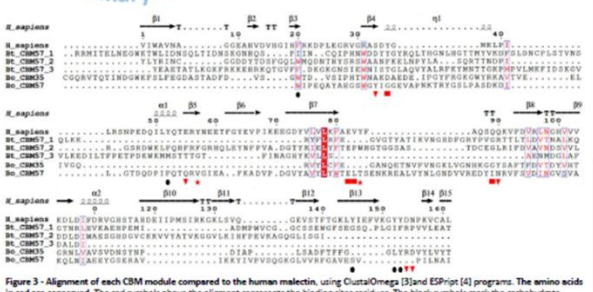


Figure 3 - **Alignment of each CBM module compared to the human malectin**, using ClustalOmega [3]and ESPript [4] programs. **The amino acids in red are conserved.** The red symbols above the alignment represents the binding sites residues. The black symbols mark the carbohydrate-interacting residues from malectin putative binding site.

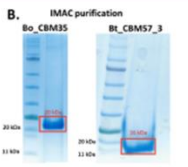
CBM modules characterization

TWO CONSTRUCTS → EXPRESSION TESTS → LARGER SCALE PROTEIN PRODUCTION → PURIFICATION → GLYCAN MICROARRAYS → X-RAY CRYSTALLOGRAPHY

1. CBM modules production:

A. Expression tests

Expressed soluble protein	Best expression condition
Bo_CBM35 (20kDa)	BL21, 37°C, 3 hours, 0.5 mM IPTG
Bt_CBM57_3 (16kDa)	Tuner, 37°C, 3 hours, 0.2 mM IPTG



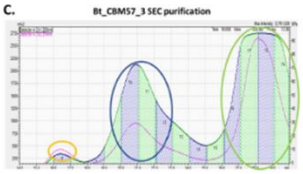
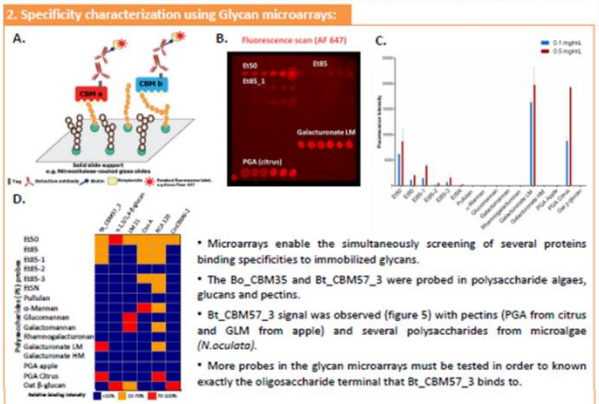


Figure 4 - **Production and purification of the proteins.** *A-* Various expression conditions were tried for soluble protein production of the five CBM modules. Two CBM modules were successfully expressed-Bo_CBM35 and Bt_CBM57_3. B- SDS-PAGE of the two CBM modules after IMAC purification. C- Size-exclusion chromatography (SEC) of Bt_CBM_57. D- SDS-PAGE of the Bt_CBM57_3 after SEC purification.

2. Specificity characterization using Glycan microarrays:



- Microarrays enable the simultaneously screening of several proteins binding specificities to immobilized glycans.
- The Bo_CBM35 and Bt_CBM57_3 were probed in polysaccharide algaes, glucans and pectins.
- Bt_CBM57_3 signal was observed (figure 5) with pectins (PGA from citrus and GLM from apple) and several polysaccharides from microalgae *(N.oculata)*.
- More probes in the glycan microarrays must be tested in order to known exactly the oligosaccharide terminal that Bt_CBM57_3 binds to.

Figure 5 – **Glycan Microarrays screening analysis.** Schematic representation of the glycan microarrays assay setup. B- Binding assay of the Bt_CBM57_3 against several polysaccharide purification fractions of the microalgae (*N.oculata*), glucans and pectins. C- Graphic of the fluorescence intensity of the Bt_CBM57_3 binding assay. Two protein concentrations were tested. D- The heat map represents the relative binding intensities calculated as the percentage of the fluorescence signal intensity given by the probe most strongly bound by each protein (normalized as 100%). Protein of known specificity were used as control: α 1,3/1,4-β-glucan and LM 21 (antibodies); Con A and RCA 120 (plant lectins) and a characterized CBM from *Cellvibrio mixtus*.

3. Structurally characterization using X-ray crystallography:

- Over 384 crystallization conditions were tested for each two proteins.
- Crystals were obtained for Bt_CBM57_3 (Figure 6), but diffraction *in house* was not observed due to protein denaturation.

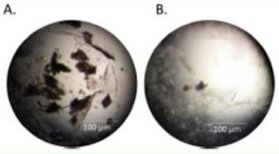


Figure 6 - **Crystallisation condition for Bt_CBM57_3** at 30 mg/ml (A) and 15 mg/ml (B) with lithium sulphate and Polyethylene glycol (PEG).

Concluding Remarks

- Both Bo_CBM35 and Bt_CBM57_3 were successfully expressed and purified for further analysis.
- Preliminary data revealed that Bt_CBM57_3 recognizes carbohydrates different from malectin ligands.
- More crystallisations trials must to be performed.
- Co-crystallisation with the ligands are in future plans.

References
[1] Schallus, Thomas, et al. "Malectin: a novel carbohydrate-binding protein of the endoplasmic reticulum and a candidate player in the early steps of protein N-glycosylation." *Molecular biology of the cell* 19.8 (2008): 3404-3414.
[2] Pinheiro, BA, *et al*, *unpublished*
[3] *The EMBL-EBI bioinformatics web and programmatic tools framework.* Nucleic acids research 43 (W1) :W580-4
[4] Robert, X. and Gouet, P. (2014) *Nucl. Acids Res.* 42(W1), W320-W324

Acknowledgements
We would like to thank Professor Manuel Coimbra, Dr. Cláudia Nunes and Carolina Panderaça, from University of Aveiro, for kindly providing polysaccharide samples.

Funding
Project supported by the Portuguese Science and Technology Foundation (FCT-MEC) through grants PTDC/BBB-BEP/0869/2014; PTDC/QUI-QUI/112537/2009; REC/BBB-BEP/0124/2012; SFRH/BPD/68563/2010; PD/BD/105727/2014 (DTQB-MolBioS Program). And by Unidade de Ciências Biomoleculares Aplicadas – UCIBIO, which is financed by national funds from FCT/MEC (UID/Multi/04378/2013) and co-financed by the ERDF under the PT2020 Partnership Agreement (POCI-01-0145-FEDER-007728).

Find us!
www.ucibio.pt

Decoding Type-C CBM modules specificities in human gut bacteria

Frederico Lourenço[1], Viviana G. Correia[1], Catarina Leote[1], Joana L.A. Brás[3], Carlos M.G.A. Fontes[2,3], Maria João Romão[1], Ana Luísa Carvalho[1], Angelina S. Palma[1], Benedita A. Pinheiro[1]

[1] UCIBIO-REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal;
[2] CIISA/FMV-UL, Faculdade de Medicina Veterinária, Universidade de Lisboa, Av. da Universidade Técnica, 1300-477 Lisboa, Portugal;
[3] NZYTech, Lda, Estrada do Paço do Lumiar, Campus do Lumiar, 1649-038 Lisboa, Portugal

Malectin and homologues

Malectin, located in the Endoplasmic reticulum, is a highly-conserved lectin in the animal kingdom. It is highly specific to the di-glucosylated oligosaccharide ($Glc_2Man_9GlcNAc_2$), an intermediate of the N-glycosylation pathway, crucial for quality control of glycoproteins folding [1].

Malectin is composed of a β-sandwich core and the binding site is located in the loop region (figure 1) [2]. Several **carbohydrate binding modules (CBMs)** were discovered across other kingdoms of life that share amino acid sequence homology with malectin, being grouped in the CBM57 family of the Carbohydrate Active Enzymes Database (CAZy).

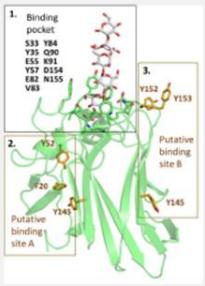


Figure 1 - Malectin structure with $Glc_2Man_9GlcNAc_2$ solved by X-ray crystallography. Identification of the Binding pocket residues as well as the putative binding residues

The CBM57 family and structurally related groups (CBM6 and CBM35) are type-C CBMs and have, like malectin, the binding pocket present in the loop region. They are expected to have diverse carbohydrate binding specificities.

Same Specificity?

PROJECT'S AIM:

- Determine their carbohydrate specificity (using glycan microarrays).
- Structurally characterize the selected novel CBM (using X-ray crystallography).

Identifying protein architecture and sequence similarity

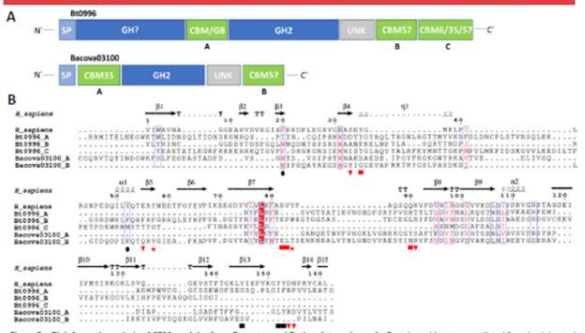


Figure 2 – Bioinformatic analysis of CBM modules from *B. ovatus* and *B. thetaiotaomicron*. A - Protein architecture predicted from both bacteria using InterproScan. Besides the CBM modules, the proteins have a signal peptide (SP), glycosidase hydrolase modules (GH) and a domain of unknown function (UNK). The selected CBM modules for the study are represented as green boxes. B - Alignment of each CBM module with human malectin, using ClustalOmega [3]and ESPript [4] programs. **The amino acids in rectangles are conserved.** The red symbols below the alignment represent the binding site carbohydrate-interacting residues. The black circle and rectangle mark the carbohydrate-interacting residues

- Two proteins containing **CBM57 and CBM35 modules from the human gut bacteria *Bacteroides ovatus* and *Bacteroides thetaiotaomicron*** were retrieved from the CAZy database. Their architecture was predicted (Figure 2A).
- **Because of the presence of several CBM modules, each CBM sequence was aligned to malectin** (figure 2B) to investigate possible different carbohydrate specificities.
- **Malectin glycan-interacting amino acid residues do not seem to be conserved.**
- The Bt0996_C showed **a higher level of divergence in the alignment**, and was selected for biochemical and structural characterization.

CBM modules characterization

TWO CONSTRUCTS → EXPRESSION TESTS → LARGER SCALE PROTEIN PRODUCTION → PURIFICATION → GLYCAN MICROARRAYS → X-RAY CRYSTALLOGRAPHY

1. Bt0996_C module production:

- Strategy used to produce Bt0996_C module for further studies (figure 4)

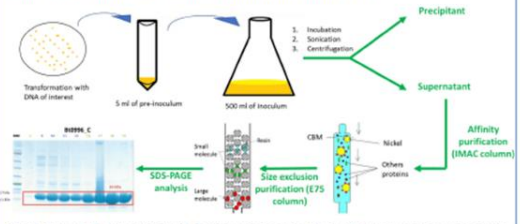
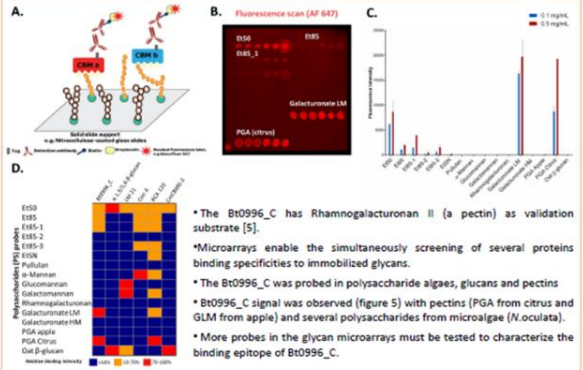


Figure 3– Schematic illustration of the Bt0996_C module production and purification. The transformed cells were inoculated in LB medium. The cells were lysed using a sonicator, followed by centrifugation. The module was purified using affinity and size exclusion chromatography. SDS-PAGE shows the result after the size exclusion purification.

2. Specificity characterization using Glycan microarrays:



- The Bt0996_C has Rhamnogalacturonan II (a pectin) as validation substrate [5].
- Microarrays enable the simultaneously screening of several proteins binding specificities to immobilized glycans.
- The Bt0996_C was probed in polysaccharide algaes, glucans and pectins
- Bt0996_C signal was observed (figure 5) with pectins (PGA from citrus and GLM from apple) and several polysaccharides from microalgae (*N.oculata*).
- More probes in the glycan microarrays must be tested to characterize the binding epitope of Bt0996_C.

Figure 5 – Glycan Microarrays screening analysis. Schematic representation of the glycan microarrays assay setup. B- Binding assay of the Bt0996_C against several polysaccharide purification fractions of the microalgae (*N.oculata*), glucans and pectins. C- Graphic of the fluorescence intensity of the Bt0996_C binding assay. Two protein concentrations were tested. D- The heat map represents the relative binding intensities calculated as the percentage of the fluorescence signal intensity given by the probe most strongly bound by each protein (normalized as 100%). Protein of known specificity were used as control: α 1,3/1,4-β-glucan and LM 21 (antibodies); Con A and RCA 120 (plant lectins) and a characterized CBM from *Cellvibrio mixtus*.

3. Structural characterization using X-ray crystallography:

- Over 384 different conditions were tested for crystallization.
- Crystals were obtained for Bt0996_C (Figure 6, diffracting at to 1.5Å resolution
- The collected data was successfully indexed. Crystals belong to the monoclinic space group C2 with cell constants a=99.8Å, b=42.8Å; c= 69.5Å; β=92.4º
- 3D structure determination is currently on going.

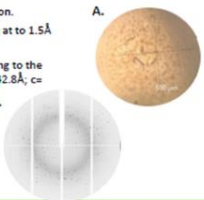


Figure 5 – A – Crystals of Bt0996_C grown with lithium sulphate and polyethylene glycol; B - X-ray diffraction pattern of the Bt0996_C crystal

Concluding Remarks

- The Bt0996_C was successfully expressed and purified for further analysis.
- Preliminary data revealed that Bt0966_C recognizes pectin polysaccharides.
- Co-crystallisation with the ligands are in future plans.

References

[1] Schallus, Thomas, et al. "Malectin: a novel carbohydrate-binding protein of the endoplasmic reticulum and a candidate player in the early steps of protein N-glycosylation." *Molecular biology of the cell* 19.8 (2008): 3404-3414.
[2] Pinheiro, BA, et al, unpublished
[3] *The EMBL-EBI bioinformatics web and programmatic tools framework*. Nucleic acids research 43 (W1): W580-4
[4] Robert, X. and Gouet, P. (2014). *Nucl. Acids Res.* 42(W1), W320-W324
[5] Martens, et al "Recognition and degradation of Plant Cell Wall Polysaccharide by Two Human Gut Symbionts" Plos Biol, 2011, 9,

Acknowledgements
We would like to thank Professor Manuel Coimbra, Dr. Cláudia Nunes and Carolina Pandeirada, from University of Aveiro, for kindly providing polysaccharide samples.

Funding
Project supported by the Portuguese Science and Technology Foundation (FCT-MEC) through grants PTDC/BBB-BEP/0969/2014, PTDC/QUI-QUI/121537/2009, REC/BBB-BEP/0124/2012, SFRH/BPD/68563/2010, PD/BD/105727/2014 (BTQB-MultiBio Program). And by Unidade de Ciências Biomoleculares Aplicadas – UCIBIO, which is financed by national funds from FCT/MEC (UID/Multi/04378/2013) and co-financed by the ERDF under the PT2020 Partnership Agreement (POCI-01-0145-FEDER-007728).

Find us!
www.ucibio.pt